1

Technical Report 681

Forecasting Device Effectiveness:
Volume III. Analytic Assessment of Device Effectiveness Forecasting Technique

Andrew M. Rose
American Institutes for Research

Anne W. Martin
Decisions and Designs, Inc.

and

Louise G. Yates
Army Research Institute

Training and Simulation Technical Area
Training Research Laboratory

U. S. Army
Research Institute for the Behavioral and Social Sciences

June 1985

U. S. ARMY RESEARCH INSTITUTE

FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the

Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON Technical Director	L. NEALE COSBY Colonel, IN Commander

Research performed under contract
for the Department of the Army

American Institutes for Research

Technical review by

Joseph D. Hagman
Michael J. Singer

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-TST, 5001 Eisenhower Avenue, Alexandria, Virginia 22333.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
ARI Technical Report 681

2. GOVT ACCESSION NO.
AD-A160 029

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
FORECASTING DEVICE EFFECTIVENESS: VOLUME III. ANALYTIC ASSESSMENT OF DEVICE EFFECTIVENESS FORECASTING TECHNIQUE

5. TYPE OF REPORT & PERIOD COVERED
Final Report: Vol. 3 of 3
August 1983-December 1984

6. PERFORMING ORG. REPORT NUMBER
--

7. AUTHOR(s)
Andrew M. Rose (AIR), Anne W. Martin (DDI), and Louise G. Yates (ARI)

8. CONTRACT OR GRANT NUMBER(s)
MDA 903-82-C-0414

9. PERFORMING ORGANIZATION NAME AND ADDRESS
American Institutes for Research
1055 Thomas Jefferson Street, NW
Washington, DC 20007

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
2Q263744A795
3350, 3.4.1

11. CONTROLLING OFFICE NAME AND ADDRESS
U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria, VA 22333-5600

12. REPORT DATE
June 1985

13. NUMBER OF PAGES
82

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)
--

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
--

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

--

18. SUPPLEMENTARY NOTES

Louise G. Yates, Contracting Officer's Representative. Volume I of Forecasting Device Effectiveness numbered TR 680; Volume II numbered RP 85-25.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Transfer of training,	Interrater agreement
Sensitivity analysis,	Monte Carlo analysis
Reliability.	

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Several analytic procedures were conducted to address various aspects of the scalar properties of the Device Effectiveness Forecasting Technique (DEFT). These procedures included Monte Carlo simulations to assess the interpretation of DEFT output, sensitivity of DEFT parameters, comparison of outputs, stability, and interrater agreement. Results indicated that it would be necessary to incorporate assumptions regarding expected distributions of input variables in order to meaningfully interpret DEFT output. Also, the Monte Carlo analyses demonstrated the sensitivity (Continued)

ARI Technical Report 681

20. (Continued)

of DEFT output scores to variations in inputs, and assessed the effects of various assumptions regarding measurement error on output scores.

The interrater agreement issue was addressed by having several raters apply DEFT to three actual training devices. Results indicated a high degree of consistency among raters for all devices for all levels of DEFT.

Keywords:

FLD 19

Technical Report 681

Forecasting Device Effectiveness:
Volume III. Analytic Assessment of Device Effectiveness Forecasting Technique

Andrew M. Rose
American Institutes for Research

Anne W. Martin
Decisions and Designs, Inc.

and

Louise G. Yates
Army Research Institute

Submitted by
Stanley F. Bolin, Acting Chief
Training and Simulation Technical Area

Approved as technically adequate
and submitted for publication by
Harold F. O'Neil, Jr., Director
Training Research Laboratory

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

June 1985

Army Project Number
2Q263744A795

Training and Simulation

Approved for public release; distribution unlimited.

ARI Research Reports and Technical Reports are intended for sponsors of R&D tasks and for other research and military agencies. Any findings ready for implementation at the time of publication are presented in the last part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

FOREWORD

Army training developers need tools to aid in the design, acquisition, and use of simulation- and computer-based programs of instruction for weapon operation and maintenance. One critical need is a job aid for the design and evaluation of training devices during all stages in the weapon acquisition cycle.

This series of three reports describes one approach to such aiding--a hybrid of decision analysis and mathematical modeling. The approach provides numerical estimates of device effectiveness which are based on expert ratings of trainee and task characteristics, functional and physical similarity between the proposed device and the operational equipment, and the instructional characteristics of the device. It is an analytic, computer-based technique--a menu-driven system--which can be used at any stage of training device design.

The product of this research can help training device procurers such as PM-TRADE and training developers in TRADOC make better documented decisions about training device design.

Edgar M Johnson

EDGAR M. JOHNSON
Technical Director

ACKNOWLEDGMENTS

The authors would like to acknowledge the assistance of the raters who contributed their time, effort, and brainpower:

George R. Wheaton
Daniel B. Felker
Harris H. Shettel
David L. Winter
Basil MacDonald.

Forecasting Device Effectiveness: III. Analytic Assessment of DEFT

EXECUTIVE SUMMARY

Requirement:

To analytically address the numeric and scalar properties of the Device Effectiveness Forecasting Technique (DEFT); to conduct an examination of interrater agreement by analyzing three training devices.

Procedure:

Several analytic procedures were conducted to address various aspects of the scalar properties of DEFT. These procedures included Monte Carlo simulations to assess the interpretation of DEFT output, sensitivity of DEFT parameters, comparison of outputs, stability, and interrater agreement.

Findings:

Results indicated that it would be necessary to encorporate assumptions regarding expected distributions of input variables in order to meaningfully interpret DEFT output. Also, the Monte Carlo analyses demonstrated the sensitivity of DEFT output scores to variations in inputs, and assessed the effects of various assumptions regarding measurement error on output scores.

The interrater agreement issue was addressed by having several raters apply DEFT to three actual training devices. Results indicated a high degree of consistency among raters for all devices and for all levels of DEFT.

Utilization of Findings:

These findings indicate that, with few modifications, DEFT can be used effectively and reliably to analytically evaluate training device-based training systems.

FORECASTING DEVICE EFFECTIVENESS: III. ANALYTIC ASSESSMENT OF DEFT

CONTENTS

LIST OF FIGURES AND TABLES

Page

FIGURES

TABLES

LIST OF FIGURES AND TABLES (Continued)

Page

TABLES (Continued)

1. Introduction

This report is submitted in partial fulfillment of Contract MDA 903-82-C-0414 between the Army Research Institute (ARI) and the American Institutes for Research (AIR). It is part of a progammatic effort to develop and analytically evaluate a model designed to forecast training device effectiveness. Specifically, this report describes the analytic evaluation phase of the effort.

Previous reports in this series have discussed issues related to the evaluation of a training system (Rose & Wheaton, 1984a), and presented an analytic model (Rose & Wheaton, 1984b). This model, named the Device Effectiveness Forecasting Technique (DEFT), incorporates numerous ratings and judgments regarding components of the training situation and the operational performance requirement and generates forecasts of training device effectiveness. In lieu of empirical tests, Rose and Wheaton (1984a) outlined several analytic methods that could be employed to assess the adequacy of such a model.

Decisions and Designs, Inc. (DDI) and AIR employed five such methods in the evaluation of DEFT:

- Interpretation of output--what sorts of results can be expected from DEFT?
- Sensitivity analysis--what is the impact on DEFT output of varying input parameter values?
- Comparison of outputs--what do differences in scores received by various devices mean?
- Stability--what is the impact of disagreement between raters on component scores?
- Interrater agreement--applying DEFT to three training devices, to what extent do raters agree for each of the various ratings and judgments?

The first four of these questions were addressed using Monte Carlo analysis. The general approach was to simulate applications of the DEFT model by generating 5,000 random values (within the appropriate ranges) for each of the various DEFT inputs (Performance Deficit, Difficulty, etc.)* and combining them according to the DEFT formulae, yielding 5,000 DEFT output scores. For the "interpretation of output" issue, this analysis, repeated under different

*For details regarding the components of DEFT, combination rules, output variables, and rating procedures, see Rose & Wheaton, 1984(b).

conditions, constituted the entire computational activity. Sensitivity analysis was performed using a variation on the basic analysis: Random values were generated for all but one of the input parameters; to examine the sensitivity of the output score to the value of the remaining input parameter, this parameter was stepped through its range of values in an orderly fashion, and output scores were computed for each of the values that it assumed. For "comparison of outputs," the basic analysis was performed twice to obtain two 5,000-element vectors of output scores. One vector was subtracted from the other, resulting in a vector of differences. A frequency distribution computed for this vector allows significance testing of difference values. Finally, the impact of less than perfect interrater stability was explored by simulating "measurement error" and scale bias and examining their effects on the DEFT output.

The basic procedure for assessing interrater agreement was to have six raters apply DEFT to three training devices. Model outputs were compared using various statistical techniques. This document presents the results of the five sets of analyses. First, we will present the general technical approach to the Monte Carlo analyses, followed by those results. We will then present the details of the interrater agreement study.

2. Monte Carlo Analyses

General Technical Approach to the Monte Carlo Analysis

As we mentioned in the introduction, Monte Carlo analysis was used to simulate applications of DEFT in order to address each of the four basic questions (interpretation, sensitivity, comparison of outputs, and stability).

* Eight input variables were used in these analyses:

- Performance Deficit (PD)
- Difficulty (D)
- Training Acquisition Efficiency (AE)
- Residual Performance Deficit (RPD)
- Residual Learning Difficulty (RLD)
- Physical Similarity (PS)
- Functional Similarity (FS)
- Transfer Efficiency (TT)

* Abbreviations are those used in report II.

These variables are obtained in different ways for each of the three levels of DEFT. However, since these different methods all result in equivalent scales (e.g., "Performance Deficit" has a range of 0-100 for all three DEFT levels), it was decided to use these variables in the Monte Carlo analyses.

Since the distribution of DEFT outputs (the basic product of each analysis) depends on the distribution of the inputs, selection of input distributions was key. Because DEFT is a new tool that has not been applied to the evaluation of a large number of training devices, no empirical distributions of inputs currently exist. Therefore, it was necessary to use artificial input distributions. The analysts working on this task selected the uniform distribution (i.e., all input values have the same probability of being selected) as the standard for input to the Monte Carlo analyses. This represents an extremely conservative approach; it was selected to provide a "worst case" baseline for comparisons with other sets of assumptions.

In addition to selecting a distributional form for input to the analyses, it was necessary to decide on the number of trials or simulated model applications for each analysis. The selection criterion used for the number of trials was the degree of convergency of (1) a distribution of data points generated randomly from an underlying uniform distribution with (2) the theoretical uniform distribution. Convergence was examined for numbers of trials ranging from 1,000 to 9,000. The number 5,000 was chosen, finally, because it is cost-effective for this application; convergence is almost as good for 5,000 trials as for 9,000 trials, and substantially less computing power is required.

Thus, each Monte Carlo analysis of DEFT output simulates 5,000 random applications of the DEFT model. This basic analysis was performed under a variety of conditions that depended upon the question to be answered. Tabular and, where appropriate, graphic presentations of results appear in the following sections.

Interpretation of Output

The objective of this first set of analyses was to explore the distributional characteristics of the DEFT output. This was done under five different conditions, three using uniform distributions, and two using truncated normal distributions. The conditions were:

1) Uniform input distributions; denominator input variables (i.e., acquisition and transfer efficiency measures [see Rose Rose & Wheaton, 1984b, Chapter 6]) range from one to 100; all others range from zero to 100. Inputs combined using initial DEFT model.

2) Uniform in ut distributions; all input variables range from one to 100. Inputs combined using initial DEFT model.

3) Uniform input distributions; all inputs range from one to 100. Square root taken of denominator (efficiency) variables (e.g., AE = $\sqrt{R/100}$ instead of AE = R/100; otherwise, combination identical to initial DEFT model.

4) Input distributions truncated normal. Inputs combined using initial DEFT model.

5) Input distributions truncated normal. Square root taken of efficiency variables. Otherwise, combination identical to initial DEFT model.

Tables 1 through 3 summarize results for intermediate and output variables under Conditions 1, 2, and 3. In these tables:

Table 1. CONDITION 1 RESULTS--UNIFORM INPUT;
INITIAL RANGES AND COMBINATIONS

DESCRIPTIVE STATISTICS FOR MODEL DEFT
5000 TRIALS

VARIABLE	MEAN	VARIANCE	STD DEV	MINIMUM	MAXIMUM
TP	24.87	491.21	22.16	.00	99.00
ACQ(A)	131.36	177317.58	421.09	.00	8722.00
AD	16.76	555.15	23.56	.00	99.00
TRP	41.71	1039.74	32.25	.00	168.22
TRANS(T)	217.69	390344.31	624.73	.00	11967.00
TOTAL(A+T)	349.04	572816.19	756.85	.00	12268.29

Table 2. CONDITION 2 RESULTS--UNIFORM INPUT;
ALL RANGES 1-100; INITIAL COMBINATION

DESCRIPTIVE STATISTICS FOR MODEL DEFT
5000 TRIALS

VARIABLE	MEAN	VARIANCE	STD DEV	MINIMUM	MAXIMUM
TP	25.12	486.31	22.05	.04	100.00
ACQ (A)	131.75	150900.78	388.46	.06	8700.00
AD	16.99	557.64	23.61	.00	97.00
TRP	42.59	1069.20	32.70	.06	188.00
TRANS (T)	211.42	398557.33	631.31	.07	11450.00
TOTAL (A+T)	343.17	555211.20	745.12	1.63	11466.21

Table 3. CONDITION 2 RESULTS--UNIFORM INPUT;
ALL RANGES 1-100; SQUARE ROOT TRANSFORMATION

DESCRIPTIVE STATISTICS FOR MODEL DEFT
5000 TRIALS

VARIABLE	MEAN	VARIANCE	STD DEV	MINIMUM	MAXIMUM
TP	25.37	488.26	22.10	.03	99.00
ACQ (A)	47.23	3751.51	61.25	.06	872.20
AD	16.58	544.03	23.32	.00	98.00
TRP	42.04	1026.01	32.03	.08	148.22
TRANS (T)	78.33	8156.73	90.31	.09	1201.60
TOTAL (A+T)	125.56	11770.24	108.49	.96	1284.90

TP = Training Problem
(A) ACQ = Total Acquisition Score
AD = Additional Deficit
TRP = Transfer Problem
(T) TRANS = Total Transfer Score
(A+T) TOTAL = Total Score.

The most striking features of these results are the high variances displayed in Conditions 1 and 2; the output distributions are extremely diffuse given uniform input distributions. In Condition 3, the output distributions are substantially tighter because of the square root transformation in the denominators (the transformation makes the denominator larger, narrowing the range).

Since the obtained values for the variance of scores in the first two conditions would make the interpretation of DEFT output relatively meaningless, we decided to modify the assumption of uniform input distributions. Based on our familiarity with training devices in general, and with U.S. Army training devices in particular, we hypothesized distributions for each input parameter. The truncated normal input distributions for Conditions 4 and 5 were the following:

	VARIABLE	MODE	RANGE
	PD (Performance Deficit)	70	30-90
	D (Difficulty)	55	10-100
	AE (R) (Training Efficiency)	65	25-100
	RPD (Residual Performance Deficit)	30	1-65
(RLD)	RD (Residual Learning Difficulty)	50	10-90
	PS (Physical Similarity)	80	30-100
	FS (Functional Similarity)	70	45-100
(TT)	RR (Transfer Efficiency)	35	10-90

These distributions were obtained by transforming a standard normal distribution centered at zero and truncated at -3 and +3. The mode of the standard normal distribution (always zero) was mapped to the mode of the target range, and the truncated value of -3 was mapped to the endpoint furthest below the mode (e.g., for a mode of 70 and a range of 30-90, -3 was mapped to 30); finally, the target distribution was truncated appropriately at the other end of the range.

Results for Conditions 4 and 5 are summarized in Tables 4 and 5. Variances are substantially lower for both of these conditions than for Conditions 1 through 3, because of the changes in the assumptions about input distributions; and variance is lower for Condition 5 than Condition 4 on account of the square root transformation.

Table 4. CONDITION 4 RESULTS--TRUNCATED NORMAL INPUT; INITIAL DEFT COMBINATIONS

DESCRIPTIVE STATISTICS FOR MODEL DEFT (RESTRICTED RANGES)
5000 TRIALS

VARIABLE	MEAN	VARIANCE	STD DEV	MINIMUM	MAXIMUM
PD	68.12	133.16	11.54	30.32	89.99
D	54.84	223.19	14.94	11.02	99.62
R (AE)	65.04	171.62	13.10	26.84	100.00
RPD	30.23	127.40	11.29	1.10	64.71
RD (RLD)	50.40	177.50	13.32	10.32	89.60
PS	76.14	187.78	13.70	30.40	99.98
FS	70.17	91.65	9.57	45.12	99.70
RR (TT)	38.53	239.35	15.47	10.01	89.54
TP	37.33	144.59	12.02	5.85	81.07
ACQ (A)	59.97	561.05	23.69	8.10	196.83
AD	10.23	127.95	11.31	.00	52.86
TRP	25.48	176.43	13.28	.83	78.95
TRANS (T)	80.49	3954.16	62.88	1.46	612.74
TOTAL (A+T)	140.46	4381.49	66.19	22.12	650.34

Table 5. CONDITION 5 RESULTS--TRUNCATED NORMAL INPUT; SQUARE ROOT TRANSFORMATION OF DENOMINATOR

DESCRIPTIVE STATISTICS FOR MODEL DEFT (RESTRICTED RANGES)
5000 TRIALS

VARIABLE	MEAN	VARIANCE	STD DEV	MINIMUM	MAXIMUM
PD	68.12	133.16	11.54	30.32	89.99
D	54.84	223.19	14.94	11.02	99.62
R (AE)	65.04	171.62	13.10	26.84	100.00
RPD	30.23	127.40	11.29	1.10	64.71
RD (RLD)	50.40	177.50	13.32	10.32	89.60
PS	76.14	187.78	13.70	30.40	99.98
FS	70.17	91.65	9.57	45.12	99.70
RR (TT)	38.53	239.35	15.47	10.01	89.54
TP	37.33	144.59	12.02	5.85	81.07
ACQ (A)	47.04	253.81	15.93	6.88	122.90
AD	10.23	127.95	11.31	.00	52.86
TRP	25.48	176.43	13.28	.83	78.05
TRANS (T)	44.07	674.29	25.97	1.10	193.88
TOTAL (A+T)	91.11	896.99	29.95	18.54	233.76

Thus, based on some reasonable assumptions regarding the distribution of expected input values, we see that DEFT outputs are interpretable and meaningful in both an absolute and a relative sense. For example, a device receiving a Training Problem (TP) score of 65.0 could be interpreted as addressing a "larger" problem than a typical device (mean = 37.33, s.d. = 12.02, Condition 4). Differences between ratings for two devices on obtained scores could be interpreted with reference to expected scores.

Sensitivity

Eight sensitivity analyses were performed, one for each of the DEFT input parameters. The objective of these analyses was to explore the impact of changes in input parameter values on the values of intermediate and output variables.

The analyses were conducted using Condition 3 of DEFT (as described above)--all input variables are assumed to be distributed uniformly between one and 100; training and transfer efficiency variables are subjected to square root transformations.

Table 6 shows DEFT results when all inputs vary freely; Tables 7 through 14 show how these results vary with systematic variation of each input parameter.

As might have been expected, the efficiency variables have the largest effect on the means and standard deviations of the output scores. For example, across the range of input values, changing training efficiency scores produces variations in the Total Score mean from 334.0 to 103.5, and changes the standard deviation from 140.0 to 96.0. In general, varying each of the other inputs changes the Total Score by approximately 100 points and the standard deviation by approximately 40 points.

Another way of looking at these results is to say that all scales (except Efficiency) have equivalent effects on the Total Score--an extreme value on any single scale will have the same effect as an extreme value on any other. Hence, all scales are "weighted" equally. The logical (and analytic) exceptions are the efficiency scales: a device that incorporates poor training or transfer principles would be expected to have a larger effect on training time, expense, and effort than any single component, since poor techniques will affect all aspects of the training and/or transfer problem.

Table 6. DESCRIPTIVE STATISTICS FOR DEFT--
FOR COMPARISON WITH SENSITIVITY ANALYSES

DESCRIPTIVE STATISTICS FOR MODEL DEFT SENSITIVITY ANALYSIS
5000 TRIALS

NAME	MEAN	VARIANCE	ST DEV	MINIMUM	MAXIMUM
PD	50.73	822.50	28.68	1.00	100.00
D	50.09	832.99	28.86	1.00	100.00
R (AE)	51.04	829.88	28.81	1.00	100.00
RPD	50.25	829.97	28.81	1.00	100.00
RD (RLD)	50.53	826.55	28.75	1.00	100.00
PS	50.17	829.67	28.80	1.00	100.00
FS	50.58	846.28	29.09	1.00	100.00
RR (TT)	50.45	834.94	28.90	1.00	100.00
TP	25.53	498.47	22.33	.03	99.00
ACQ (A)	47.08	3654.26	60.45	.03	837.20
AD	16.79	559.71	23.66	.00	98.00
TRP	42.00	1035.41	32.18	.02	179.14
TRANS (T)	79.74	9694.16	98.46	.03	1211.60
TOTAL (A+T)	126.83	13410.99	115.81	1.16	1266.40

Table 7. SENSITIVITY ANALYSIS FOR PD

SENSITIVITY ANALYSIS FOR PD

VARIABLE:	TP		ACQ (A)		AD		TRP		TRANS (T)		TOTAL (A+T)	
PD	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV
1	.5	.3	.9	.9	16.5	23.7	42.0	32.4	77.9	93.6	78.8	93.6
25	12.7	7.2	23.6	23.0	16.5	23.7	42.0	32.4	77.9	93.6	101.4	96.1
50	25.4	14.4	47.2	46.0	16.5	23.7	42.0	32.4	77.9	93.6	125.0	103.8
75	38.1	21.6	70.7	69.0	16.5	23.7	42.0	32.4	77.9	93.6	148.6	115.6
100	50.8	28.8	94.3	92.0	16.5	23.7	42.0	32.4	77.9	93.6	172.2	130.5

Table 8. SENSITIVITY ANALYSIS FOR D

SENSITIVITY ANALYSIS FOR D

VARIABLE:	TP		ACQ (A)		AD		TRP		TRANS (T)		TOTAL (A+T)	
D	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV
1	.5	.3	.9	.9	16.5	23.7	42.0	32.4	77.9	93.6	78.8	93.6
25	12.6	7.1	23.3	22.6	16.5	23.7	42.0	32.4	77.9	93.6	101.2	96.1
50	25.2	14.2	46.6	45.2	16.5	23.7	42.0	32.4	77.9	93.6	124.5	103.6
75	37.7	21.3	70.0	67.7	16.5	23.7	42.0	32.4	77.9	93.6	147.8	115.1
100	50.3	28.4	93.3	90.3	16.5	23.7	42.0	32.4	77.9	93.6	171.1	129.6

Table 9. SENSITIVITY ANALYSIS FOR R (AE)

SENSITIVITY ANALYSIS FOR R (AE)

VARIABLE:	TP		ACQ (A)		AD		TRP		TRANS (T)		TOTAL (A+T)	
R	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV
1	25.6	22.2	256.2	221.9	16.5	23.7	42.0	32.4	77.9	93.6	331.0	240.0
25	25.6	22.2	51.2	44.4	16.5	23.7	42.0	32.4	77.9	93.6	129.1	103.2
50	25.6	22.2	36.2	31.4	16.5	23.7	42.0	32.4	77.9	93.6	114.1	98.1
75	25.6	22.2	29.6	25.6	16.5	23.7	42.0	32.4	77.9	93.6	107.4	96.8
100	25.6	22.2	25.6	22.2	16.5	23.7	42.0	32.4	77.9	93.6	103.5	96.0

Table 10. SENSITIVITY ANALYSIS FOR RPD (PD')

SENSITIVITY ANALYSIS FOR RPD

VARIABLE:	TP		ACQ (A)		AD		TRP		TRANS (T)		TOTAL (A+T)	
RPD	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV
1	25.6	22.2	47.7	61.3	16.5	23.7	17.0	23.7	31.8	60.9	79.5	85.4
25	25.6	22.2	47.7	61.3	16.5	23.7	29.2	24.8	51.3	69.7	102.0	92.1
50	25.6	22.2	47.7	61.3	16.5	23.7	41.8	27.8	77.7	84.3	125.4	103.8
75	25.6	22.2	47.7	61.3	16.5	23.7	54.4	32.2	101.1	102.0	148.8	118.9
100	25.6	22.2	47.7	61.3	16.5	23.7	67.1	37.5	124.5	121.6	172.1	134.2

Table 11. SENSITIVITY ANALYSIS FOR RD (D') (RLD)

SENSITIVITY ANALYSIS FOR RD (RLD)

VARIABLE:	TP		ACQ (A)		AD		TRP		TRANS (T)		TOTAL (A+T)	
RD	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV
1	25.6	22.2	47.7	61.3	16.5	23.7	17.0	23.7	31.8	60.9	79.5	85.4
25	25.6	22.2	47.7	61.3	16.5	23.7	29.1	24.7	54.4	71.3	102.0	93.4
50	25.6	22.2	47.7	61.3	16.5	23.7	41.7	27.6	77.8	87.1	125.5	[illegible]
75	25.6	22.2	47.7	61.3	16.5	23.7	54.3	31.9	101.2	105.8	148.9	[illegible]
100	25.6	22.2	47.7	61.3	16.5	23.7	66.9	37.1	124.7	126.1	172.4	[illegible]

Table 12. SENSITIVITY ANALYSIS FOR PS

SENSITIVITY ANALYSIS FOR PS

VARIABLE:	TP		ACQ (A)		AD		TRP		TRANS (T)		TOTAL (A+T)	
PS	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV
1	25.6	22.2	47.7	61.3	.0	.0	25.4	22.2	46.9	59.4	91.6	84.2
25	25.6	22.2	47.7	61.3	3.1	6.3	28.5	23.1	52.8	61.1	100.5	[illegible]
50	25.6	22.2	47.7	61.3	12.4	16.0	37.8	27.4	70.4	80.5	118.1	101.5
75	25.6	22.2	47.7	61.3	27.9	24.8	53.3	33.3	97.4	104.6	147.0	121.1
100	25.6	22.2	47.7	61.3	49.6	29.0	75.1	36.6	140.2	130.9	187.8	144.4

Table 13. SENSITIVITY ANALYSIS FOR FS

SENSITIVITY ANALYSIS FOR FS

VARIABLE:	TP		ACQ (A)		AD		TRP		TRANS (T)		TOTAL (A+T)	
FS	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV
1	25.6	22.2	47.7	61.3	48.5	29.1	74.0	36.6	137.6	130.8	185.3	[illegible]
25	25.6	22.2	47.7	61.3	27.8	24.9	53.2	33.4	97.0	105.3	146.7	[illegible]
50	25.6	22.2	47.7	61.3	12.4	16.1	37.8	27.4	70.1	80.6	117.8	101.8
75	25.6	22.2	47.7	61.3	3.1	6.5	28.5	23.1	52.8	63.8	100.5	[illegible]
100	25.6	22.2	47.7	61.3	.0	.0	25.4	22.2	46.9	59.4	91.6	84.2

Table 14. SENSITIVITY ANALYSIS FOR RR (R') (TT)

SENSITIVITY ANALYSIS FOR RR (TT)

VARIABLE:	TF		ACQ (A)		AD		TRF		TRANS (T)		TOTAL (A+T)	
RR	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV
1	25.6	22.2	47.7	61.3	16.5	23.7	42.0	32.4	119.5	321.3	167.2	[illegible]
25	25.6	22.2	47.7	61.3	16.5	23.7	42.0	32.4	83.9	64.9	131.6	88.4
50	25.6	22.2	47.7	61.3	16.5	23.7	42.0	32.4	59.3	45.9	107.0	75.6
75	25.6	22.2	47.7	61.3	16.5	23.7	42.0	32.4	48.4	37.4	96.1	71.0
100	25.6	22.2	47.7	61.3	16.5	23.7	42.0	32.4	42.0	32.4	89.6	68.6

Comparison of Outputs

The objective of the "comparison of outputs" analysis is to determine the probability of any given level of difference between two DEFT TOTAL scores. To this end, two DEFT TOTAL output vectors (5000 data points each) were generated, and one was subtracted from the other to obtain a frequency distribution of differences. Table 15 summarizes the three distributions.

It should be noted that the two TOTAL distributions were generated using Condition 3 above, which assumes uniformly distributed inputs; as was noted before, this is an extremely conservative assumption.

Figure 1 shows a frequency distribution of the differences; as is to be expected, the differences are distributed approximately normally with a mean very close to zero.

Table 16 summarizes the probability distribution based on this analysis. This table can be used to determine statistical significance, although it is extremely conservative due to the underlying distributional assumptions. According to this table, two devices would need to differ by approximately 150 points in the Total Score to be judged

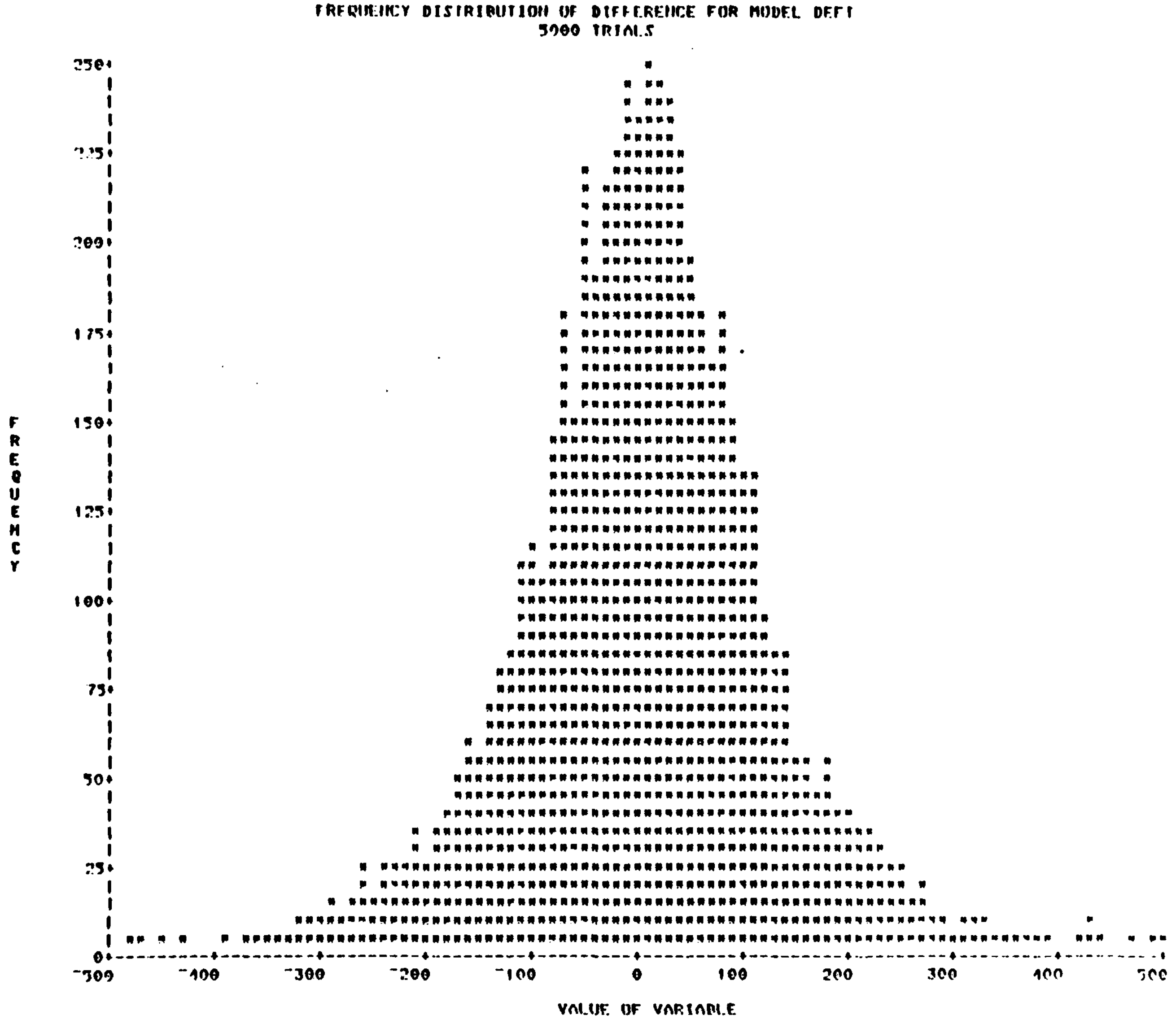
Table 15. DESCRIPTIVE STATISTICS FOR DEFT CONDITION 3
DIFFERENCE ANALYSIS

```
DESCRIPTIVE STATISTICS FOR MODEL DEFT  DIFFERENCE ANALYSIS
                         5000 TRIALS

NAME          MEAN    VARIANCE    ST DEV   MINIMUM   MAXIMUM
----          ----    --------    -- ---   -------   -------
TOTAL1      126.52    12869.87    113.45       .87   1335.32

TOTAL2      126.51    12320.60    111.00      1.56   1163.96

DIFFER         .01    24404.78    156.22  ¯1118.48   1222.42
```



NOTE. EACH * REPRESENTS ABOUT 5 DATA POINT(S)

Figure 1. FREQUENCY DISTRIBUTION OF DEFT CONDITION 3 DIFFERENCES

Table 16. PROBABILITY DISTRIBUTION OF DEFT CONDITION 3 DIFFERENCES

SIGNIFICANCE TABLE FOR MODEL DEFT DIFFERENCE ANALYSIS
5000 TRIALS

DIFF	CUM %	DIFF	CUM %	DIFF	CUM %	DIFF	CUM %	DIFF	CUM %
¯1140	.0000	¯660	.0044	¯180	.0752	300	.9734	780	.9986
¯1130	.0000	¯650	.0044	¯170	.0834	310	.9750	790	.9986
¯1120	.0000	¯640	.0044	¯160	.0932	320	.9766	800	.9986
¯1110	.0002	¯630	.0046	¯150	.1024	330	.9784	810	.9988
¯1100	.0002	¯620	.0046	¯140	.1144	340	.9794	820	.9990
¯1090	.0002	¯610	.0048	¯130	.1280	350	.9806	830	.9990
¯1080	.0002	¯600	.0052	¯120	.1422	360	.9818	840	.9990
¯1070	.0002	¯590	.0054	¯110	.1602	370	.9828	850	.9990
¯1060	.0002	¯580	.0054	¯100	.1792	380	.9834	860	.9990
¯1050	.0002	¯570	.0062	¯90	.1968	390	.9840	870	.9990
¯1040	.0002	¯560	.0068	¯80	.2212	400	.9844	880	.9990
¯1030	.0004	¯550	.0068	¯70	.2512	410	.9848	890	.9990
¯1020	.0006	¯540	.0070	¯60	.2764	420	.9856	900	.9992
¯1010	.0008	¯530	.0074	¯50	.3132	430	.9876	910	.9992
¯1000	.0008	¯520	.0080	¯40	.3450	440	.9882	920	.9992
¯990	.0008	¯510	.0086	¯30	.3808	450	.9886	930	.9994
¯980	.0008	¯500	.0090	¯20	.4180	460	.9888	940	.9996
¯970	.0008	¯490	.0094	¯10	.4582	470	.9894	950	.9996
¯960	.0008	¯480	.0100	0	.4970	480	.9896	960	.9996
¯950	.0008	¯470	.0106	10	.5384	490	.9904	970	.9996
¯940	.0010	¯460	.0110	20	.5790	500	.9910	980	.9996
¯930	.0010	¯450	.0118	30	.6186	510	.9914	990	.9998
¯920	.0010	¯440	.0122	40	.6562	520	.9920	1000	.9998
¯910	.0010	¯430	.0132	50	.6886	530	.9920	1010	.9998
¯900	.0012	¯420	.0134	60	.7184	540	.9922	1020	.9998
¯890	.0014	¯410	.0136	70	.7460	550	.9926	1030	.9998
¯880	.0014	¯400	.0140	80	.7756	560	.9926	1040	.9998
¯870	.0016	¯390	.0150	90	.8004	570	.9932	1050	.9998
¯860	.0016	¯380	.0154	100	.8226	580	.9936	1060	.9998
¯850	.0016	¯370	.0164	110	.8418	590	.9936	1070	.9998
¯840	.0016	¯360	.0176	120	.8608	600	.9940	1080	.9998
¯830	.0016	¯350	.0182	130	.8752	610	.9942	1090	.9998
¯820	.0016	¯340	.0192	140	.8896	620	.9944	1100	.9998
¯810	.0016	¯330	.0198	150	.8990	630	.9948	1110	.9998
¯800	.0018	¯320	.0218	160	.9084	640	.9954	1120	.9998
¯790	.0020	¯310	.0236	170	.9156	650	.9956	1130	.9998
¯780	.0020	¯300	.0250	180	.9246	660	.9958	1140	.9998
¯770	.0020	¯290	.0276	190	.9312	670	.9958	1150	.9998
¯760	.0022	¯280	.0296	200	.9380	680	.9962	1160	.9998
¯750	.0026	¯270	.0322	210	.9434	690	.9962	1170	.9998
¯740	.0026	¯260	.0364	220	.9494	700	.9968	1180	.9998
¯730	.0028	¯250	.0392	230	.9546	710	.9972	1190	.9998
¯720	.0032	¯240	.0436	240	.9586	720	.9976	1200	.9998
¯710	.0032	¯230	.0480	250	.9624	730	.9980	1210	.9998
¯700	.0034	¯220	.0520	260	.9652	740	.9980	1220	.9998
¯690	.0038	¯210	.0582	270	.9682	750	.9984	1230	1.0000
¯680	.0040	¯200	.0626	280	.9702	760	.9984	1240	1.0000
¯670	.0042	¯190	.0684	290	.9722	770	.9986	1250	1.0000

as "different" at the 0.10 probability level. Much more realistic is a difference based on the restricted ranges generated in Conditions 4 and 5, described earlier. In these cases, for example, a difference of 30 points in the Total Score (Condition 5) would make two devices a standard deviation apart.

Stability Analyses

The purpose of the stability analyses was to examine the impact of deviations from perfect reliability. It is normally assumed that a rather high degree of stability is necessary to demonstrate the validity of the measuring instrument and/or the robustness of the effect being measured. Establishing the existence of the desired degree of stability is an empirical endeavor (e.g., through repeated observations of raters); nonetheless, Monte Carlo analyses can be used to hypothetically examine the potential impact of instability.

Two kinds of Monte Carlo analyses were performed. The scale bias analysis shows the impact of preferences for certain portions of the input scale. The two-judge random error analysis examines the effect of measurement error on apparent stability. Results of this analysis can be used for null hypothesis testing.

Impact of scale bias. Table 17 summarizes the results of the scale bias analysis, which investigates the impact of a rater's preference for any specific portion cf the allowable 1-100 scale. Inputs are assumed to be uniformly distributed; each row in Table 17 represents a different range from which the values for all input variables are drawn. The first row, provided for comparison, shows intermediate and output variable results for the unbiased case, in which the entire 1-100 range is used. Subsequent rows show results for cases in which simulated judgments (i.e., input values) are confined to smaller portions of the scale.

Two-judge random error analysis. As has already been mentioned, Monte Carlo analysis cannot be used to determine the degree of stability; this is an empirical question. However, investigation can be made of the impact of measurement error on apparent stability. In particular, suppose that two judges are in agreement about all aspects of a device, but, due to measurement error, their ratings do not coincide perfectly. How does this affect their apparent agreement?

To investigate this question, five sets of simulated DEFT model output were generated. The first set represents the "truth" in the form of 5,000 random applications of

Table 17. SCALE BIAS ANALYSIS FOR DEFT (UNIFORM INPUT DISTRIBUTIONS)

SCALE BIAS ANALYSIS FOR MODEL DEFT
5000 TRIALS

VARIABLE:	TP		ACQ (A)		AD		TRP		TRANS (T)		TOTAL (A+T)	
SCALE	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV	MEAN	ST DEV
-----	----	-- ---	----	-- ---	----	-- ---	----	-- ---	----	-- ---	----	-- ---
1-100	25.3	22.3	46.2	58.2	16.5	23.2	42.2	31.7	77.1	88.3	123.3	105.4
1- 20	1.1	.9	4.0	3.9	3.2	4.5	4.3	4.5	15.7	19.6	19.7	20.0
20- 40	9.0	2.5	16.6	4.9	3.3	4.7	12.4	5.3	22.9	10.1	39.5	11.2
40- 60	25.0	4.1	35.5	6.2	3.3	4.7	28.4	6.2	40.3	9.1	75.8	11.0
60- 80	49.0	5.8	58.6	7.4	3.3	4.7	52.4	7.4	62.8	9.2	121.4	11.7
80-100	80.9	7.4	85.4	8.3	3.3	4.7	84.4	8.7	89.1	9.6	174.5	12.7
20- 80	24.9	12.8	37.0	20.7	10.0	14.1	35.1	18.7	52.4	30.6	89.4	36.8
1- 25	1.7	1.4	5.6	5.6	4.0	5.6	5.7	5.8	19.1	23.4	24.7	24.0
25- 50	14.0	3.9	23.2	6.9	4.2	5.9	18.3	7.0	30.3	12.1	53.5	13.8
50- 75	39.0	6.5	49.6	8.7	4.2	5.9	43.3	8.6	55.0	11.5	104.6	14.4
75-100	76.5	9.0	81.9	10.3	4.2	5.9	80.8	10.7	86.6	12.0	168.6	15.7
25- 75	24.9	10.5	36.4	16.6	8.3	11.7	33.4	15.6	49.0	24.5	85.3	29.5
1- 33	2.9	2.4	8.5	9.0	5.3	7.5	8.2	7.8	24.6	29.4	33.1	30.7
33- 67	24.9	7.1	35.8	10.8	5.7	8.0	30.7	10.5	44.1	16.0	79.9	19.2
67-100	69.6	11.4	76.5	13.3	5.5	7.7	75.3	13.6	82.8	15.8	159.4	20.6
1- 50	6.5	5.6	16.0	18.1	8.2	11.5	14.7	12.6	36.7	42.4	52.6	46.0
50-100	56.1	15.6	65.7	19.5	8.3	11.7	64.7	19.3	75.8	24.0	141.5	30.8

DEFT in which two judges in fact agree perfectly on each and every input value. Table 18 summarizes this set of output (generated under Condition 3). The other four sets of DEFT output represent various kinds of "imperfection" in the form of deviation about the "truth" values. Tables 19 through 22 summarize DEFT results for hypothetical judges whose ratings (input values) vary randomly about the "true" value.

In Tables 19 and 20, the random variation is uniform over the interval true value ±5 (interval width 10); in Tables 21 and 22, the variation is uniform over the interval true value ±10 (interval width 20).

Table 23 summarizes distributions of difference in DEFT TOTAL among the various data sets. The first row (DIF10J1X) describes the variation of hypothetical Judge 1's DEFT TOTAL about "truth's" DEFT TOTAL when Judge 1 is assumed to be reliable to ±5; the second row (DIF10J2X) summarizes the same variation for hypothetical Judge 2. The third row (DIF10J1J2) summarizes the distribution of differences between Judge 1 and Judge 2's DEFT TOTALS when the two judges are assumed to be in perfect agreement, and each is reliable to ±5. The fourth through sixth rows repeat the first through third rows for hypothetical judges that are reliable to ±10 (interval width 20).

Table 18. HYPOTHETICAL "TRUE" RESULTS FOR DEFT

DESCRIPTIVE STATISTICS FOR MODEL DEFT INTER-RATER ANALYSIS -- TRUE VALUE
5000 TRIALS

NAME	MEAN	VARIANCE	ST DEV	MINIMUM	MAXIMUM
PD	51.28	841.17	29.00	1.00	100.00
D	50.25	817.00	28.58	1.00	100.00
R (AE)	50.78	818.28	28.61	1.00	100.00
RPD	51.64	832.52	28.85	1.00	100.00
RD (RLD)	49.77	832.99	28.86	1.00	100.00
PS	50.54	821.20	28.66	1.00	100.00
FS	50.24	821.77	28.67	1.00	100.00
RR (TT)	50.80	826.48	28.75	1.00	100.00
TP	25.81	498.07	22.32	.03	98.00
ACQ (A)	48.23	3952.76	62.87	.04	809.90
AD	16.75	558.34	23.63	.00	98.00
TRP	42.50	1043.59	32.30	.06	175.49
TRANS (T)	78.00	8163.48	90.35	.06	1187.50
TOTAL (A+T)	126.23	12143.31	110.20	.98	1240.11

Table 19. RESULTS FOR HYPOTHETICAL JUDGE 1-- DEVIATION OF ± 5 FROM "TRUTH"

DESCRIPTIVE STATISTICS FOR MODEL DEFT INTER-RATER

5000 TRIALS ANALYSIS -- JUDGE #1 (INT WIDTH 10)

NAME	MEAN	VARIANCE	ST DEV	MINIMUM	MAXIMUM
PD	51.41	838.02	28.95	1.00	100.00
D	50.31	811.60	28.49	1.00	100.00
R (AE)	50.90	809.99	28.46	1.00	100.00
RPD	51.71	825.46	28.73	1.00	100.00
RD (RLD)	49.84	822.94	28.69	1.00	100.00
PS	50.61	815.46	28.56	1.00	100.00
FS	50.30	815.63	28.56	1.00	100.00
RR (TT)	50.93	822.00	28.67	1.00	100.00
TP	25.93	496.69	22.29	.03	99.00
ACQ (A)	46.83	3161.89	56.23	.03	725.40
AD	16.66	552.63	23.51	.00	97.00
TRP	42.50	1035.41	32.18	.03	177.04
TRANS (T)	76.27	6686.59	81.77	.07	949.00
TOTAL (A+T)	123.10	9796.67	98.98	1.55	1088.63

Table 20. RESULTS FOR HYPOTHETICAL JUDGE 2-- DEVIATION OF ± 5 FROM "TRUTH"

DESCRIPTIVE STATISTICS FOR MODEL DEFT INTER-RATER ANALYSIS --
JUDGE #2 (INT WIDTH 10) 5000 TRIALS

NAME	MEAN	VARIANCE	ST DEV	MINIMUM	MAXIMUM
PD	51.31	832.14	28.85	1.00	100.00
D	50.29	812.63	28.51	1.00	100.00
R (AE)	50.80	813.37	28.52	1.00	100.00
RPD	51.64	828.28	28.78	1.00	100.00
RD (RLD)	49.86	829.39	28.80	1.00	100.00
PS	50.68	814.12	28.53	1.00	100.00
FS	50.35	816.74	28.58	1.00	100.00
RR (TT)	50.87	823.44	28.70	1.00	100.00
TP	25.90	494.84	22.24	.06	100.00
ACQ (A)	47.27	3221.30	56.76	.08	601.60
AD	16.70	557.85	23.62	.00	97.00
TRP	42.50	1039.92	32.25	.07	175.65
TRANS (T)	76.53	6848.36	82.75	.07	1052.80
TOTAL (A+T)	123.80	10010.21	100.05	1.40	1107.82

Table 21. RESULTS FOR HYPOTHETICAL JUDGE 1--
DEVIATION OF ± 10 FROM "TRUTH"

DESCRIPTIVE STATISTICS FOR MODEL DEFT INTER-RATER ANALYSIS --
JUDGE #1 (INT WIDTH 20) 5000 TRIALS

NAME	MEAN	VARIANCE	ST DEV	MINIMUM	MAXIMUM
PD	51.33	825.24	28.73	1.00	100.00
D	50.17	793.37	28.17	1.00	100.00
R (AE)	50.88	806.41	28.40	1.00	100.00
RPD	51.77	820.94	28.65	1.00	100.00
RD (RLD)	49.84	818.94	28.62	1.00	100.00
PS	50.70	808.54	28.43	1.00	100.00
FS	50.37	807.97	28.42	1.00	100.00
RR (TT)	50.84	805.47	28.38	1.00	100.00
TP	25.81	479.47	21.90	.01	98.01
ACQ (A)	46.15	2764.10	52.57	.02	558.00
AD	16.63	551.31	23.48	.00	97.00
TRP	42.49	1027.32	32.05	.02	187.05
TRANS (T)	74.97	6282.88	79.26	.03	985.60
TOTAL(A+T)	121.12	9132.09	95.56	1.65	988.53

Table 22. RESULTS FOR HYPOTHETICAL JUDGE 2--
DEVIATION OF ± 10 FROM "TRUTH"

DESCRIPTIVE STATISTICS FOR MODEL DEFT INTER-RATER ANALYSIS --
JUDGE #2 (INT WIDTH 20) 5000 TRIALS

NAME	MEAN	VARIANCE	ST DEV	MINIMUM	MAXIMUM
PD	51.36	824.22	28.71	1.00	100.00
D	50.34	800.38	28.29	1.00	100.00
R (AE)	50.95	801.64	28.31	1.00	100.00
RPD	51.72	819.54	28.63	1.00	100.00
RD (RLD)	50.00	809.37	28.45	1.00	100.00
PS	50.64	806.58	28.40	1.00	100.00
FS	50.32	802.96	28.34	1.00	100.00
RR (TT)	50.88	806.74	28.40	1.00	100.00
TP	25.95	491.39	22.17	.06	99.00
ACQ (A)	46.25	2778.98	52.72	.08	631.45
AD	16.62	541.59	23.27	.00	95.00
TRP	42.51	1009.59	31.77	.06	185.20
TRANS (T)	75.06	5967.60	77.25	.06	901.20
TOTAL (A+T)	121.31	8693.74	93.24	1.26	907.58

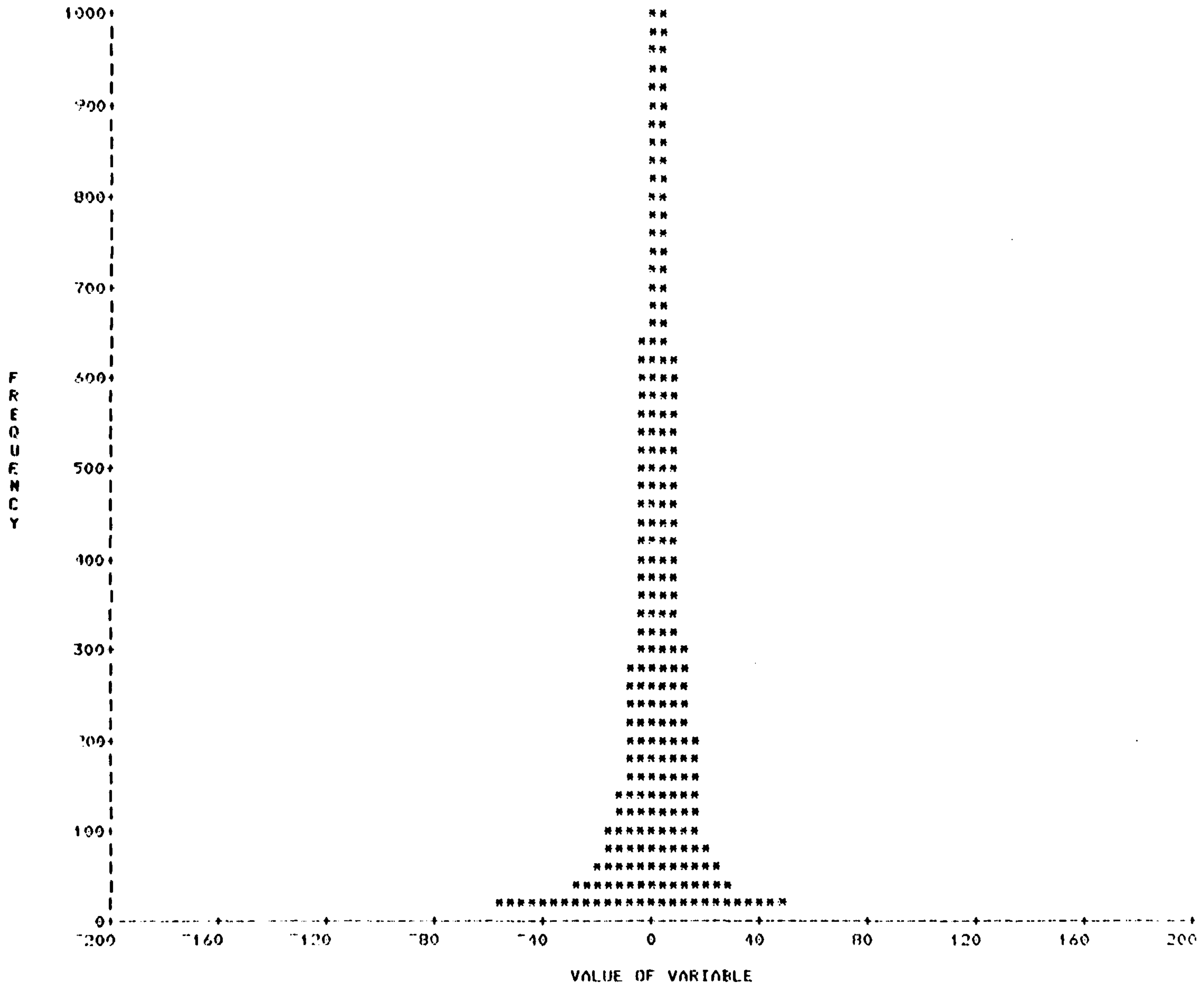
Table 23. DISTRIBUTIONS OF DEFT TOTAL DIFFERENCES

DESCRIPTIVE STATISTICS FOR MODEL DEFT INTER-RATER DIFFERENCES
5000 TRIALS

NAME	MEAN	VARIANCE	ST DEV	MINIMUM	MAXIMUM
DIF10J1X	¯3.13	1689.11	41.10	¯646.40	533.59
DIF10J2X	¯2.43	1620.27	40.25	¯623.61	458.34
DIF10J1J2	¯.70	1633.79	40.42	¯551.94	581.99
DIF20J1X	¯5.11	3529.54	59.41	¯755.65	586.10
DIF20J2X	¯4.92	3024.65	55.00	¯818.98	644.14
DIF20J1J2	¯.19	3144.11	56.07	¯641.92	703.01

The utility of this analysis is in its potential for null hypothesis testing. Given two (real) judges rating the same device, and a difference between their DEFT TOTAL scores, we can determine the likelihood of a difference of that magnitude or larger given stability of ±5 or ±10 and an assumption of no underlying disagreement. Since the differences appear to be distributed normally (see Figures 2 through 7), this test can be made using the standard normal distribution. Output of this analysis can also be used to determine confidence intervals or credible intervals about the DEFT TOTAL computed from one (real) judge's input ratings, assuming stability of ±5 or ±10.

FREQUENCY DISTRIBUTION OF JUDGE #1 - TRUE VALUE FOR INT WIDTH 10
5000 TRIALS



NOTE EACH * REPRESENTS ABOUT 20 DATA POINT(S)

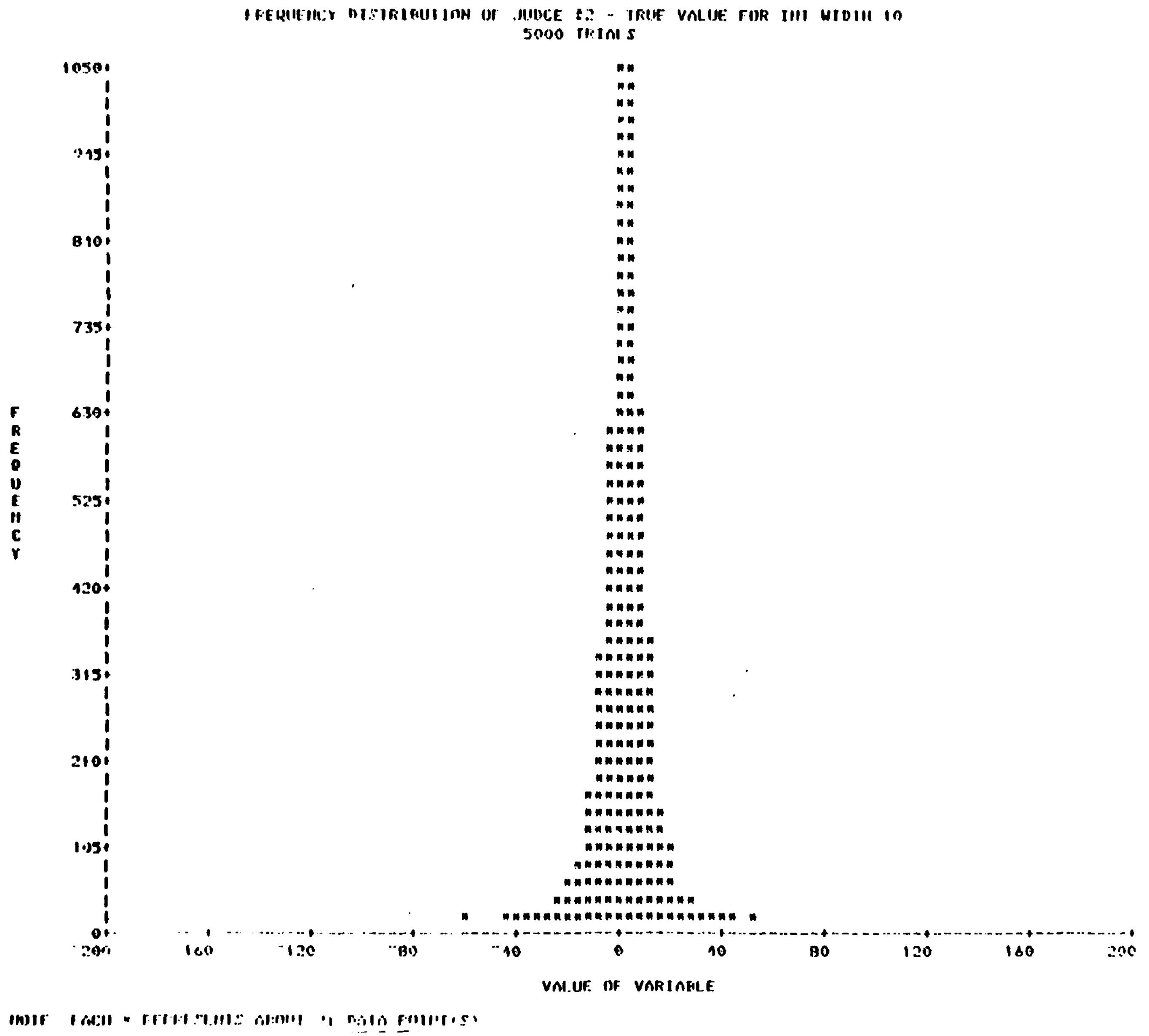
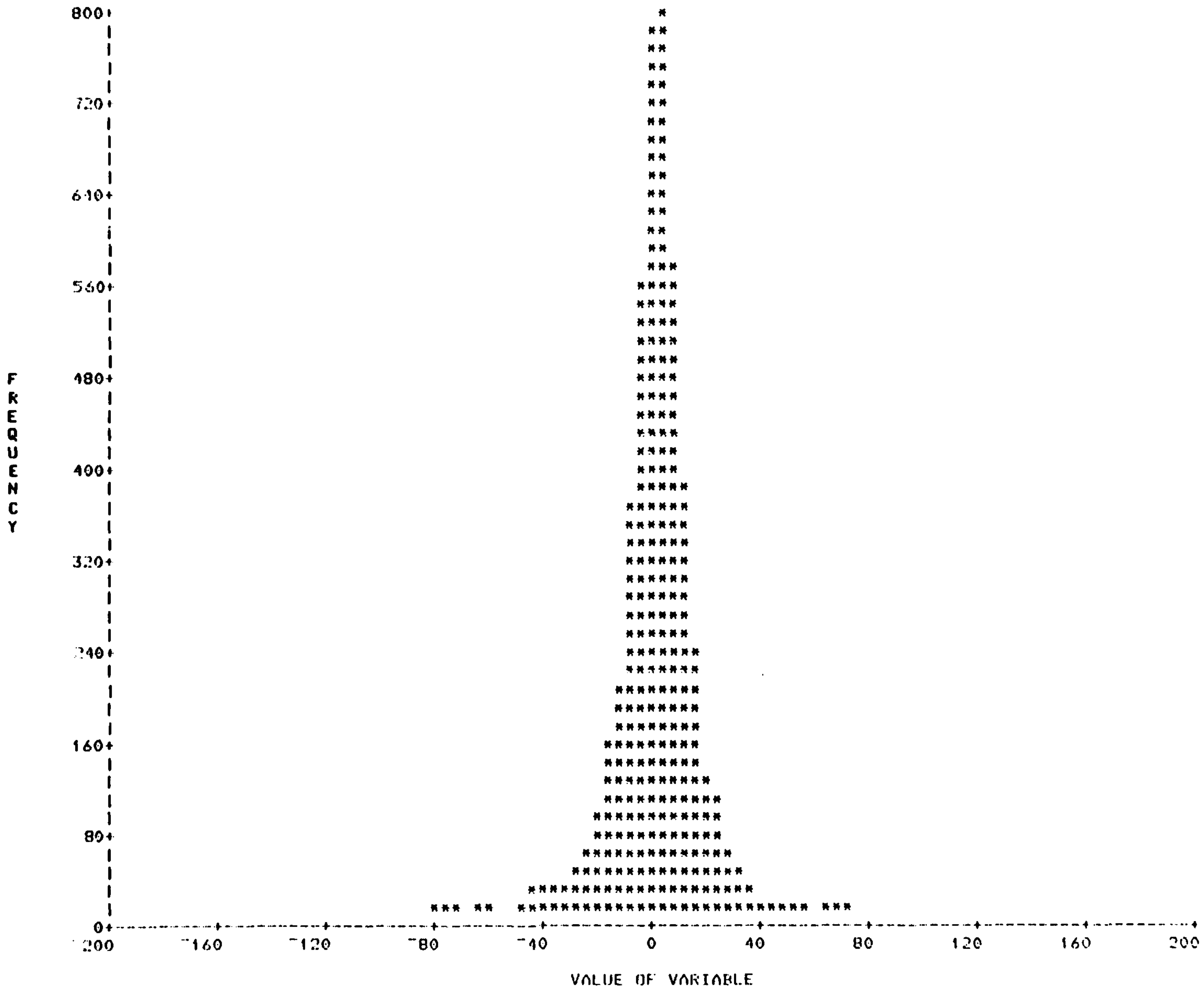


Figure 3. DISTRIBUTION OF DEFT DIFFERENCES FOR HYPOTHETICAL JUDGE 2 (RELIABLE TO ± 5) VERSUS "TRUTH"

FREQUENCY DISTRIBUTION OF JUDGE #1 - JUDGE #2 FOR INT WIDTH 10
5000 TRIALS



NOTE EACH * REPRESENTS ABOUT 16 DATA POINT(S)

Figure 4. DISTRIBUTION OF DEFT TOTAL DIFFERENCES FOR HYPOTHETICAL JUDGE 1 VERSUS HYPOTHETICAL JUDGE 2 (PERFECT

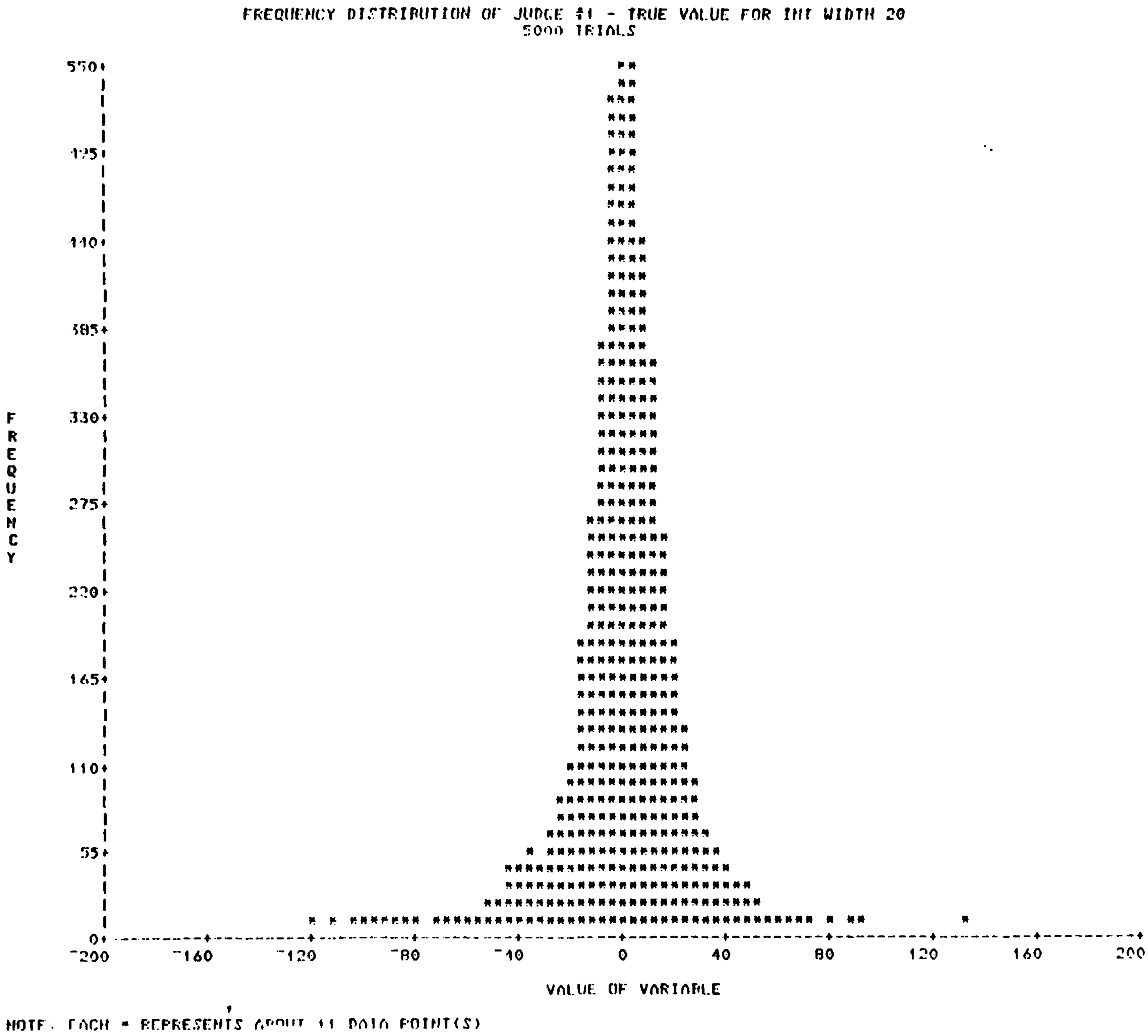
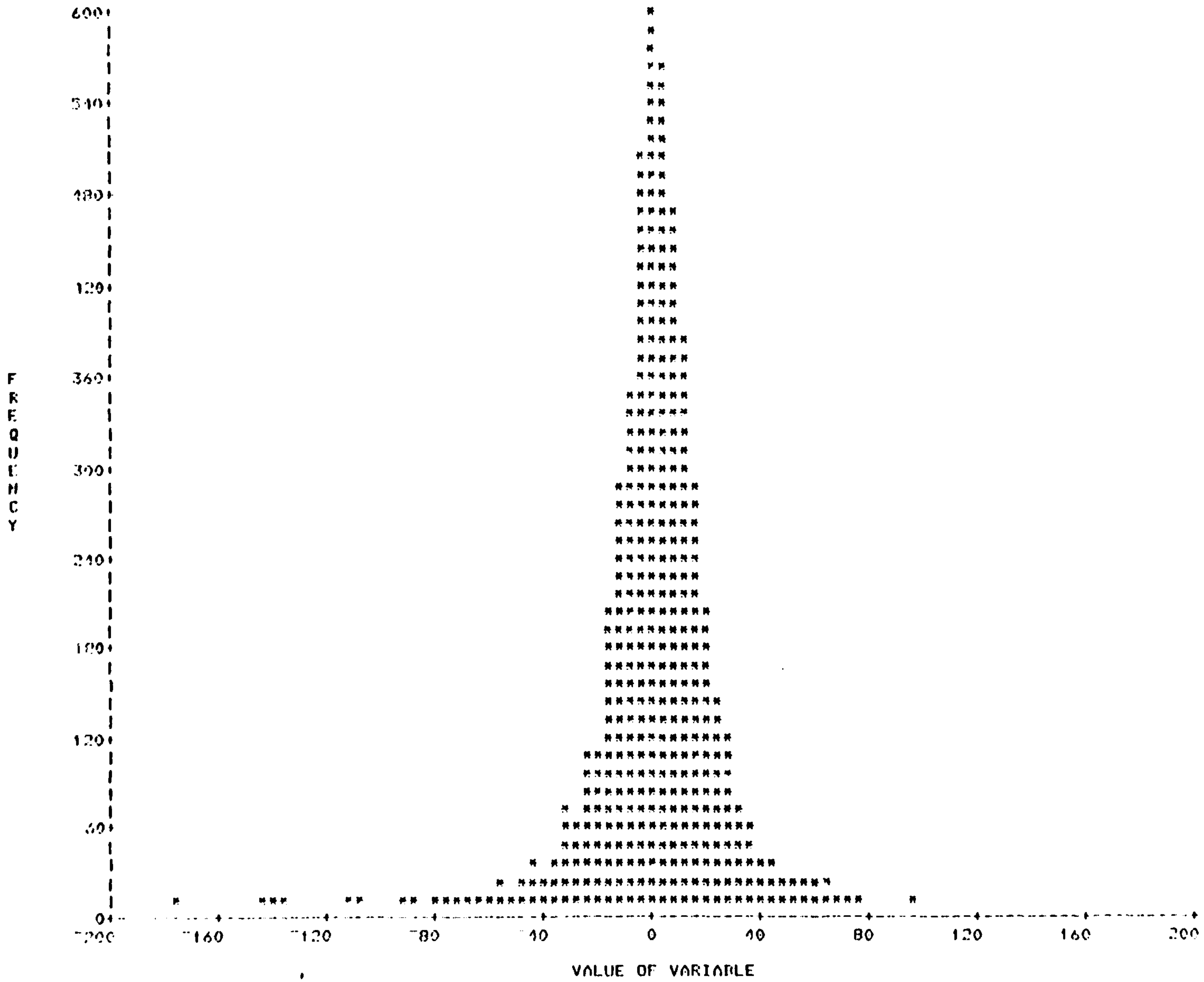


Figure 5. DISTRIBUTION OF DEFT TOTAL DIFFERENCES FOR HYPOTHETICAL JUDGE 1 (RELIABLE TO ± 10) VERSUS "TRUTH"

FREQUENCY DISTRIBUTION OF JUDGE #2 - TRUE VALUE FOR INT WIDTH 20
5000 TRIALS



NOTE EACH * REPRESENTS ABOUT 12 DATA POINT(S)

Figure 6. DISTRIBUTION OF DEFT TOTAL DIFFERENCES FOR

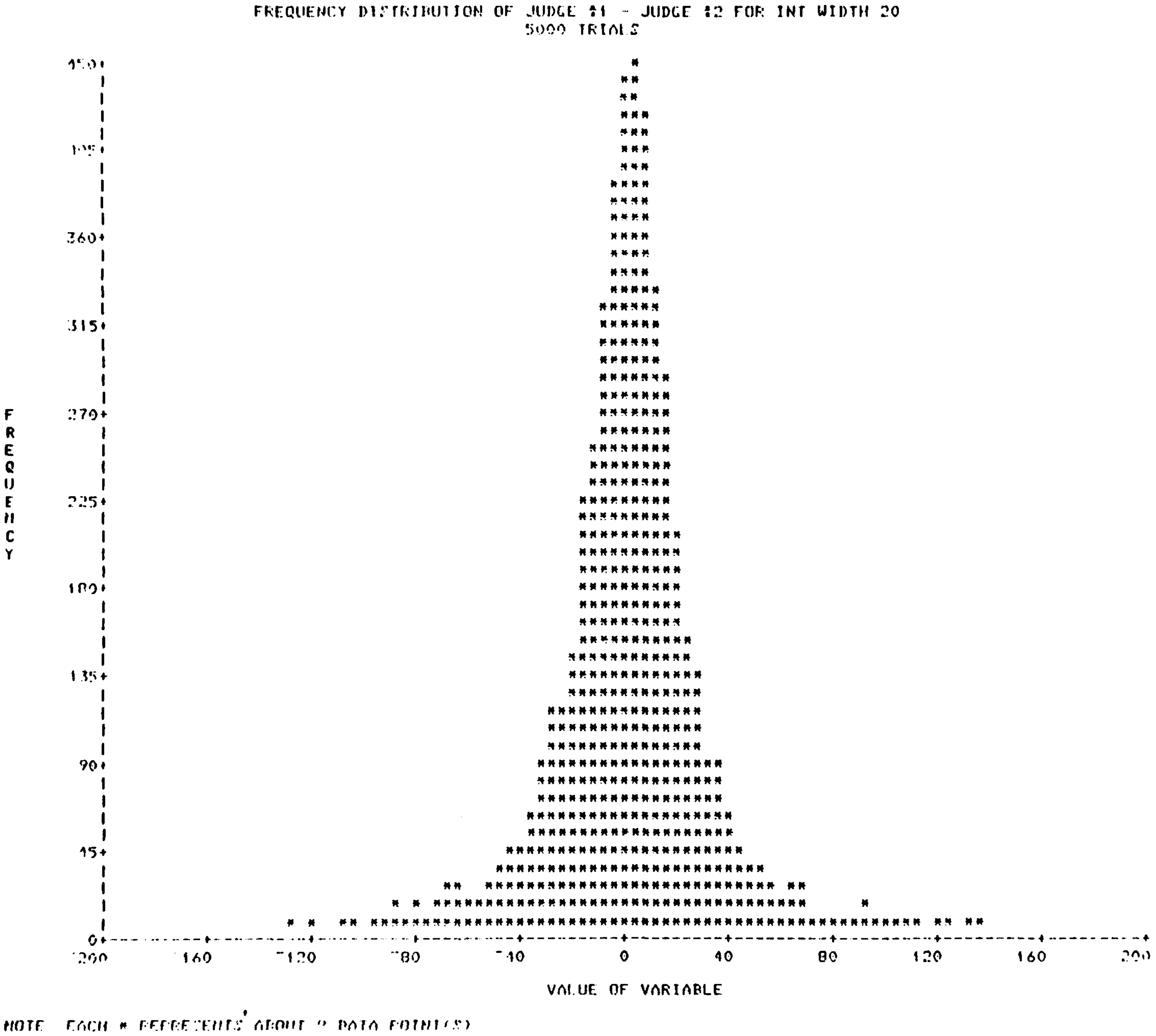


Figure 7. DISTRIBUTION OF DEFT TOTAL DIFFERENCES FOR HYPOTHETICAL JUDGE 1 VERSUS HYPOTHETICAL JUDGE 2

3. Interrater Agreement

The purpose of this exercise was to determine the degree of interrater agreement that could be achieved using DEFT. This exercise also served as a "dry run" through the DEFT procedures--in essence, a "feasibility" study. Could DEFT be used by various types of raters with more or less familiarity with the selected training devices and more or less familiarity with DEFT?

The method chosen was to have six raters use DEFT to evaluate three training devices. Two of the training devices were designed to train the same tasks and subtasks--thus, we had a "comparative" evaluation. The third training device was designed to train several different tasks. We selected two of these tasks. We chose this method -- i.e., a limited set of training devices and a limited set of raters -- rather than alternative approaches (e.g., many raters-one training device, few raters-many training devices, many raters-many training devices) primarily because of time and resource constraints. However, we also viewed this method as a"worst-case" test: if we could not demonstrate agreement in this situation, we would not be able to demonstrate agreement in less controlled situations. Our method also constrained the use of sophisticated statistical evaluations. For example, correlations between raters over repeated measures on the same rating scale could not be meaningfully interpreted due to the small number of observations. Nonetheless, descriptive statistics, such as mean differences across raters, could provide sufficient information to determine the feasibility and usefulness of DEFT.

Method

Devices and Tasks/Subtasks. Two armor gunnery training devices were selected: The MK-60 Gunnery Trainer (VIGS), and the burst-on-target (BOT) trainer. These two devices were examined in the context of training a single gunnery engagement, shown in Figure 8 (from Harris, Ford, Tufano, & Wiggs, 1983). The third device selected was a maintenance procedures simulator. This was selected because AIR staff were intimately involved in its design, extensive materials were available, and the tasks selected for evaluation were similar to maintenance procedures contained in U.S. Army tasks. Brief descriptions of the three devices and the tasks and subtasks evaluated follow.

IDOC JOB OBJECTIVE 56
PLUS BOT

Precision, periscope, stationary firing tank, moving tank target (1200-1600) meters), SABOT, direct fire adjustment (BOT)

GUNNER BEHAVIORAL ELEMENTS

1. Gunner indexes ammunition.
2. Gunner turns on main gun switch.
3. Gunner announces IDENTIFIED.
4. Gunner applies lead in direction of target apparent motion.
5. Gunner lays crosshair leadline at center of target vulnerability.
6. Gunner makes final precise lay.
7. Gunner announces ON THE WAY.
8. Gunner fires main gun.
9. Gunner announces sensing and BOT.
10. Gunner relays (BOT).
11. Gunner announces ON THE WAY (BOT).
12. Gunner fires main gun (BOT).

The gunnery engagement and gunner behaviors come from two sources.

1. Boldovici, J.A. (HumRRO), Boycan, G.G. (ARI), Fingerman, P.F., & Wheaton, G.R. (AIR). M601A0S Tank Gunnery Data Handbook, ARI Technical Report TR-79-A7, March 1979.

2. U.S. Army, FM17-12, Tank Gunnery, March 1977.

FIGURE 8. GUNNERY ENGAGEMENT

The gunnery engagement selected (Figure 8) was selected for several reasons. First, AIR staff were familiar with it; second, excellent documentation was available, and third, this engagement had previously been processed through earlier versions of the TRAINVICE models (see Harris et al., 1983).

The MPS Trainer. Materials drawn from AIR/Bedford files for the period 1974-1983 were extracted and edited to describe the E-3A Navigation Computer System (NCS) and the Maintenance Procedure Simulator (MPS) for that system. The MPS was built by Honeywell to E-3A design specifications developed by AIR/Bedford.

The MPS was designed and acquired to support training in organizational (flightline) maintenance procedures for the AN/ASN-118 NCS installed on the E-3A aircraft. The NCS supplies navigation data to the aircraft flight control system, the flight crew, and the radar data processing group. The NCS incorporates a pair of redundant CAROUSEL-type inertial navigation units, a single doppler system to measure altitude, and an Omega VLF receiver/computer system to measure aircraft position. Organizational maintenance of the NCS relies primarily upon automatic fault detection and isolation performed by built-in test equipment (BITE). Isolated faults are corrected by removal and replacement of line-replaceable units (LRUs) or substitutions of faulty soldered components (inductors, capacitors, filters).

The MPS is a computer-controlled trainer housed in a single integrated console. Operation of the E-3A aircraft AN/ASN-118 Navigation Computer System (NCS) is simulated only to the extent required for performance of the required organization-level maintenance procedures for the NCS. Faults in the NCS are simulated through the action of computer software. Required maintenance actions such as removal and replacement, connect and disconnect, and inspection are simulated by the use of MPS controls rather than actual operations.

During a normal training situation, the student operates controls of simulated aircraft and support equipment contained in the MPS. The computer software repetitively samples MPS control settings and causes the appropriate response to be displayed. Software response to the instructor/student actions can cause one or more of the following to occur:

1) change to one or more indicator displays

2) removal or change of 35-mm slide displays

3) Teleprinter message

The MPS provides 273 training exercises that are used to train entering E-3A maintenance technicians on the NCS system-specific operations and maintenance procedures. Students entering the training course have completed basic training and a general navigation course which leads to the award of semi-skilled (3-level) rating in AFSC 328X4. Upon graduation, students proceed to the E-3A Wing at Tinker AFB, where they begin work on the flightline. They are under supervision and receive additional on-the-job training.

Table 24 describes two "tasks" which are, in reality, two parts of one of the 273 exercises. The tasks selected for description are: (1) Checkout of the Inertial Navigation System (INS), and (2) Fault isolation of Fault 10 (of 100). Two information packages were prepared. The first set represented each task as performed in conjunction with the operational equipment. The second set represented the same tasks as performed in conjunction with MPS. Both provide data formulated for direct entry into the computerized DEFT program. The data included descriptive

Table 24. MPS and E-3A Tasks and Subtasks

Task 1: Checkout of Inertial Navigation System (INS)

Subtask Number	**Subtask Description**
10	Ensure E-3A aircraft power and cooling is available
20	Turn NCS Power on
30	Turn Autopilot off
40	Turn (2) probe heaters off
50	Synchronize (2) Horizontal Situation Indicators (HSI)
60	Set INS-1 and INS-2 to align mode
70	Test CDU displays and lamps
80	Detect Fault 10. (Performance index does not decrease from 9 to 5)
Task 2:	**Fault Isolation of Fault 10**
81	Interchange CDU-1 and CDU-2 (Simulated on MPS)
82	Perform Checkout (Task 1: 10-80)
83	Interchange INU-1 and INU-2 (Simulated on MPS)
84	Perform Checkout (Task 1: 10-80)
85	Check 115 VAC Power
86	Check wiring continuity (resistance)
87	Replace shorted capcititor (Simulated on MSP)
88	Perform Checkout (Task 1: 10-80)

text for each subtask and the controls, displays, skills, and knowledge associated with the subtask. Task 1 was detailed only to the level required to link Task 1 and Task 2. The details of the subtasks were greatly abbreviated to reduce or eliminate redundancy of activities which are required by the actual procedures, both for the operational equipment and for the trainer. Photographs and accompanying text were provided to indicate location of equipment; a listing of the associated displays and controls was also provided.

Raters. Six AIR staff members participated in this study. These raters had differing degrees of familiarity with each of the training devices, tasks, and DEFT itself:

Raters 1 and 2:	Very familiar with DEFT, BOT; familiar with VIGS; unfamiliar with MPS
Raters 3 and 4:	Unfamiliar with DEFT, BOT, and VIGS; very familiar with MPS
Raters 5 and 6:	Familiar with DEFT, BOT, and VIGS; unfamiliar with MPS.

We planned to examine the impact of these differences on the various DEFT ratings and outputs.

Procedure. Packages of materials were prepared for each training device. The packages varied in the quality and quantity of information provided. Thus, the BOT "packages" consisted of a picture of the device, a brief engineering description, and the list of tasks and subtasks involved. The VIGS package was the actual device user's manual, complete with pictures, instructions for use, and capabilities of the device. The MPS package contained scores of pictures, descriptions, engineering specifications, extracts from the Technical Manual used by actual crewmen on the E-3A aircraft, and the user's manual for MPS.

Following the distribution of these packages to each of the raters, Raters 1-5 met to discuss the packages and to receive instruction on how to use DEFT. It was decided that the sparse information available regarding the BOT device would be inadequate for the purposes of this study. (Although in a "real-world" application, training device evaluators might be faced with similar problems -- i.e., a lack of detailed information -- our primary purpose was to determine interrater agreement. If each rater supplied his own set of assumptions regarding, e.g., training proficiency standards, differences in ratings could not be attributed to disagreements regarding DEFT.) Thus, the raters were briefed as to the details of BOT, both as performed on the training devices (BOT and VIGS) and as performed on the M60 tank. In addition, raters were briefed in detail on the E-3A and MPS configurations for the tasks under investigation.

DEFT was presented and discussed at the "mechanical" level; that is, raters were told how to operate the computer and how to proceed through the DEFT analyses. There was no discussion as to the meaning or interpretation of the various judgments and scales; we hoped that the information provided on the screen would be sufficient.

Following this meeting, each rater was given a DEFT program diskette and a data diskette, containing the necessary data bases. Each rater then processed each of the three training devices through all three levels of DEFT. Raters analyzed BOT first, VIGS second, and MPS third, completing all DEFT analyses on each device before analyzing the next device.

At the completion of these analyses, the data diskettes were collected and the raw data scanned. A cursory examination of these data revealed that the information contained on the DEFT screens and the briefings held prior to the analyses were inadequate. Examination of the notes each rater kept regarding his ratings indicated that each was operating under a different set of assumptions. These differences ranged from data entry conventions (e.g., if a Training Principle in the Acquisition Efficiency analyses of DEFT III was judged to be "not applicable," some raters entered "0," others entered "100," and others entered "999") to different assumptions regarding trainee characteristics (e.g., some raters thought the trainees for the MPS device were skilled maintenance crewmen, while others thought that they were naive crewmen, while others thought that they were naive graduates of a Technical School, with no aircraft experience). Thus, it was decided to reconvene the raters to discuss the devices and clarify assumptions. Following these discussions, changes in ratings were re-entered by the individual raters. Because of logistic constraints, Raters 5 and 6 could not attend this meeting; therefore, their results were not included in further analyses.

Results

Output indexes. At each level of DEFT, seven output indexes are computed for a training device evaluation (although different numbers and types of ratings are involved in the different DEFT levels). These seven are:

1) Training Problem (TP)

2) Acquisition Efficiency (AE)

3) Acquisition (A); computed as TP/AE

4) Transfer Problem (TRP)

5) Transfer Efficiency (TE)

6) Transfer (T); computed as TRP/TE

7) Total Score; computed as A + T

Theoretically, these indexes should be equivalent across all three levels of DEFT for a particular training device evaluation, since the successively more detailed levels of DEFT are designed to be componential assessments of more global judgments. Thus, the first question we will examine is whether raters were "internally consistent": For each index on each training device, do the scores for the different levels of DEFT agree?

Relevant data are shown in Tables 25 - 27. Table 25 shows obtained indexes for each rater on the BOT device for all levels of DEFT; Table 26 shows the same information for the VIGS device; and Table 27 shows the same information for the MPS device. Note that these data were obtained after the second meeting of the raters, where assumptions involved and interpretations of the scales were discussed.

TABLE 25. DEFT INDEX VALUES: BOT

	Rater 1			Rater 2			Rater 3			Rater 4		
	DEFT 1	DEFT II	DEFT III	DEFT I	DEFT II	DEFT III	DEFT I	DEFT II	DEFT III	DEFT I	DEFT II	DEFT III
Training Problem	20.0	37.5	19.7	26.0	35.0	17.0	30.0	40.0	15.9	25.0	35.0	17.7
Acquisition Efficiency	0.80	0.85	0.67	0.67	0.81	0.48	0.70	0.81	0.46	0.80	0.76	0.66
Acquisition	25.0	44.1	29.4	38.8	43.2	35.4	42.9	49.4	34.6	31.5	46.1	26.8
Transfer Problem	22.7	55.0	27.4	21.0	62.5	33.1	28.0	57.5	16.0	26.2	70.0	33.8
Transfer Efficiency	0.25	0.42	0.30	0.25	0.50	0.21	0.40	0.40	0.33	0.35	0.45	0.10
Transfer	91.0	130.9	91.3	84.0	125.0	157.0	70.0	143.7	50.9	75.0	155.6	333.0
TOTAL	116.0	175.0	120.7	122.8	168.2	192.4	112.9	193.1	85.5	106.5	201.7	359.8

TABLE 26. DEFT INDEX VALUES: VIGS

	Rater 1			Rater 2			Rater 3			Rater 4		
	DEFT 1	DEFT II	DEFT III	DEFT I	DEFT II	DEFT III	DEFT I	DEFT II	DEFT III	DEFT I	DEFT II	DEFT III
Training Problem	17.5	42.5	22.9	24.0	50.0	23.5	25.0	55.0	19.0	20.0	45.0	20.8
Acquisition Efficiency	0.90	0.78	0.70	0.85	0.75	0.60	0.80	0.83	0.42	0.80	0.88	0.79
Acquisition	19.4	54.5	32.7	28.2	66.7	39.2	31.2	66.3	45.2	25.0	51.1	26.3
Transfer Problem	6.0	35.0	10.4	6.6	29.0	9.2	10.0	47.5	10.0	15.0	50.0	18.2
Transfer Efficiency	0.50	0.77	0.51	0.70	0.92	0.54	0.60	0.77	0.46	0.50	0.85	0.39
Transfer	12.0	45.5	20.4	9.4	31.5	17.0	16.7	61.7	21.7	30.0	58.8	46.7
TOTAL	31.4	100.0	53.1	37.6	98.2	56.2	47.9	128.0	66.9	55.0	109.9	73.0

TABLE 27. DEFT INDEX VALUES: MPS

	Rater 1			Rater 2			Rater 3			Rater 4		
	DEFT 1	DEFT II	DEFT III	DEFT I	DEFT II	DEFT III	DEFT I	DEFT II	DEFT III	DEFT I	DEFT II	DEFT III
Training Problem	18.0	25.0	5.6	16.7	30.0	6.3	24.0	40.0	6.7	16.0	22.5	9.5
Acquisition Efficiency	0.80	0.80	0.89	0.60	0.76	0.47	0.70	0.80	0.61	0.55	0.83	0.81
Acquisition	22.5	31.2	6.3	27.9	39.5	13.4	34.3	50.0	11.1	29.1	27.1	11.7
Transfer Problem	9.0	20.0	7.7	6.0	35.0	7.2	14.0	30.0	15.0	6.0	27.5	7.3
Transfer Efficiency	0.40	0.43	0.50	0.50	0.46	0.24	0.60	0.57	0.40	0.65	0.47	0.38
Transfer	22.5	46.5	15.5	12.0	76.1	30.2	23.3	52.6	37.5	9.2	58.5	19.2
TOTAL	45.0	77.7	21.8	39.9	115.6	43.6	57.6	102.6	48.6	38.3	85.6	30.9

The logical question to ask first is what an "acceptable" level of internal consistency would be. How close to one another should we desire that these indexes be? This is an arbitrary decision; however, considering the results of the Monte Carlo analyses discussed in previous sections, it is clear that the data shown in these tables for DEFT I and DEFT III are internally consistent. Of the 84 pairs (3 devices x 4 raters x 7 indexes) of DEFT I and DEFT III indexes, 70 (83.3%) are within 20 points of each other,and about half are within 10 points of each other. Furthermore, most of the large disagreements are due to arithmetic combinations of smaller disagreements. For example, consider Rater 2, BOT:

	DEFT I	DEFT III
TRP	21.0	33.1
TE	0.25	0.21
T	84.0	157.0
Total Score	122.8	192.4

The relatively small difference in TRP is magnified by the very small difference in TE to produce large differences in T and Total Score. This also may have been anticipated from the Monte Carlo sensitivity analyses: small differences in the Efficiency indexes will have large effects on summary indexes. If these cumulative differences are taken into account, it appears that DEFT I and DEFT III indexes are internally consistent.

On the other hand, DEFT II indexes are substantially higher than either DEFT I or DEFT III in practically all cases. A closer examination of the data reveals that the problems seem to be with the TP and TRP indexes (the Training and Transfer Problems, respectively). Each is approximately twice as large for DEFT II than for the others.

This anomaly can be explained by examining how these indexes are derived for DEFT II as compared to DEFT I and DEFT III. In both of the latter cases, TP and TRP are multiplicative functions of two ratings: Performance Deficit and Performance Difficulty. Thus, in DEFT I, if a training device objective is judged to contain 50% skills and knowledge not possessed by trainees, and these skills and knowledge are judged to be moderately difficult to learn -- e.g., they are rated at "50" on the Performance Difficulty scale -- the TP score will be (50 x 50)/100 = 25. However, in DEFT II, the judgment made as to the Performance Deficit is a simple "yes" or "no" (can do or can't do) for each task contained in the training objective. Thus, the multiplicative combination of deficit and difficulty is not contained in DEFT II. In fact, when the DEFT II indexes are modified by encorporating either DEFT I or DEFT III Performance Deficit ratings, the DEFT II indexes dovetail precisely with the other indexes. (These recalculated indexes are not shown.)

The other relatively minor inconsistencies in these data are in the Efficiency indexes (AE and TE) of DEFT III. In most cases (19 out of 24), the DEFT III Efficiency indexes are the lowest of the three (although in most cases these differences are quite small). In post-rating discussions, the raters felt that this was partially due to an "oversegmentation" problem: many of the eleven Training Efficiency and eight Transfer Efficiency principles received quite low ratings when applied to subtasks. For example, augmenting feedback for a relatively trivial subtask such as "Indexes ammunition" would quite reasonably <u>not</u> be included as an instructional feature of the VIGS device; nevertheless, VIGS was "penalized" with a low Efficiency rating for this principle.

Part of this problem is a terminological artifact of the particular tasks and subtasks selected for this study. While we termed "Indexes ammunition" a subtask, in standard task analyses it would probably be considered a "step" or a "behavioral element." The resolution of the Efficiency index problem will involve either "tightening up" DEFT input requirements (e.g., by specifying task-analytic procedures and definitions for determining "tasks" and "subtasks"), or by conducting DEFT III Efficiency analyses at the task level.

The next question that can be addressed by the examination of these data is interrater agreement within and across devices for these indexes. Thus, for example, do raters agree on the TP value for VIGS? Again, the question as to what would constitute "agreement" must be arbitrarily answered. Standard correlational techniques are not meaningfully interpretable with small sample sizes. Thus, we will examine interrater agreement descriptively.

When one closely examines Tables 25 - 27, one can only be impressed by the equivalence of the indexes across raters for all three training devices. With the exception of the Total Score and an occasional "deviant" point, all indexes are within a few point of one another. Considering the range of values that these indexes can take and the expected magnitude of difference scores as demonstrated by the Monte Carlo analyses, this correspondence is excellent. If the 100-point scales were converted to discrete 5- or

7-point scales, interrater agreement would be almost perfect.

Again, we must note that these data were obtained following a discussion among the raters; this discussion undoubtedly pulled the ratings closer together. (Countering this, however, is that discussions were of the rating _scales_, not of the summary indexes.) The picture of interrater agreement prior to the discussion, while still quite good, was not quite so rosy. As was mentioned previously, differing interpretations and rating conventions (particularly with respect to scoring rules for the Efficiency scales) resulted in many index values that were not comparable. For example, when a Training Principle was judged as "not applicable," some raters scored the scale as "zero," others as "100," and others as "999." Clearly, it would not make sense to compare indexes derived for these different raters.

The major discrepancy in these comparisons is the disagreements in the Total Scores. Paralleling the above discussions, we attribute these differences to the cumulative effects of smaller differences in individual component indexes; furthermore, many of the Total Score differences can be traced to the large impacts of the Efficiency indexes.

One possible solution, as suggested by the Monte Carlo analyses, is to transform the Efficiency indexes (e.g., by using a square root). While this reduces the problem, it does not eliminate it; however, this manipulation, plus the adoption of the suggestion to conduct DEFT III Efficiency analyses at the task (rather than the subtask) level, would produce significant convergence in Total Scores.

In summary, these data indicate substantial interrater agreement for all DEFT indexes and across the three devices. This is even more encouraging when one considers first that the raters had different degrees of familiarity with DEFT and the three devices, and second that the three devices were of quite different sorts. The next issue to examine is whether these levels of interrater agreements are maintained when the individual scales are examined.

Individual scales. Table 28 shows the average pairwise agreements among the four raters for each of the eight DEFT I scales. These figures were computed by taking the absolute differences between each pair of raters on each scale judgment, adding them, and calculating a mean and standard deviation. Since all raters rated all dimensions, there were six differences that were combined for each entry in the table. In addition, row and column means of these mean differences are shown.

TABLE 28. MEANS AND STANDARD DEVIATIONS OF PAIRED RATER COMPARISONS FOR EACH TRAINING DEVICE - DEFT1

					Question					
		PD	LD	TA	RD	RLD	PS	FS	TT	
Device										
BOT	$\bar{x}$	11.67	0.0	8.17	5.00	5.00	5.33	9.17	9.17	6.69
	δ	(6.88)	(0.0)	(4.95)	(2.89)	(2.89)	(2.63)	(5.34)	(5.34)	(1.80)
E3A	$\bar{x}$	12.17	5.83	14.17	5.00	9.17	10.00	12.50	14.17	10.38
	δ	(7.06)	(3.44)	(6.72)	(5.00)	(5.34)	(7.07)	(9.47)	(6.72)	(3.02)
VIGS	$\bar{x}$	9.17	10.00	5.83	5.00	12.83	13.33	10.00	11.67	9.73
	δ	(5.34)	(5.77)	(3.44)	(5.00)	(8.15)	(6.24)	(7.07)	(6.87)	(2.58)
										GRAND $\bar{x}$
	$\bar{x}$	11.0	5.28	9.39	5.00	9.00	9.56	10.56	11.67	8.93

As could be surmised from the discussion above concerning the output indexes, interrater agreement for each of the underlying scales was also quite substantial. Overall, the average disagreement was approximately 9 points (on a hundred-point scale), well within what could be considered acceptable levels of agreement. For the individual scales, the average disagreement was between 5.0 and 11.67 points, with no particular scale having an unusually high level of disagreement. Likewise, the three devices all showed equivalent levels of agreement.

Tables 29 and 30 show the equivalent data for DEFT II and DEFT III. Again, with minor discrepancies, interrater agreement was high for all scales for the DEFT models on all three devices. The conclusions to draw from these tables are the same as were made above for the summary indexes: Interrater agreement for DEFT is encouragingly high, especially given differences among raters with respect to familiarity with DEFT and the three devices; and the level of interrater agreement demonstrated would support the continued development and use of DEFT for the evaluation of training-device-based training systems.

TABLE 29. MEANS AND STANDARD DEVIATIONS OF PAIRED RATER COMPARISONS FOR EACH TRAINING DEVICE - DEFT II

						Question			
			PD	LD	RD	RLD	PS	FS	Mean
Device									
BOT	Task1	$\bar{x}$	0.0	10.83	0.0	11.67	10.83	7.50	6.81
		δ	(0.0)	(5.34)	(0.0)	(6.87)	(5.34)	(4.79)	
	Task2	$\bar{x}$	0.0	5.0	0.0	5.0	5.0	13.33	4.72
		δ	(0.0)	(5.0)	(0.0)	(2.89)	(5.0)	(9.43)	
E3A	Task1	$\bar{x}$	0.0	8.33	0.0	2.50	10.0	10.0	5.14
		δ	(0.0)	(5.53)	(0.0)	(2.50)	(7.07)	(5.77)	
	Task2	$\bar{x}$	0.0	12.50	0.0	10.0	11.67	18.33	8.75
		δ	(0.0)	(7.50)	(0.0)	(7.07)	(6.87)	(10.67)	
VIGS	Task1	$\bar{x}$	0.0	9.33	0.0	15.0	11.67	16.67	8.78
		δ	(0.0)	(5.31)	(0.0)	(8.66)	(6.87)	(7.45)	
	Task2	$\bar{x}$	0.0	10.17	0.0	10.17	12.50	15.00	7.97
		δ	(0.0)	(5.87)	(0.0)	(5.27)	(7.50)	(7.64)	
	$\bar{x}$		0.0	9.36	0.0	9.06	10.28	13.47	7.03
	δ		(0.0)	(2.56)	(0.0)	(4.55)	(2.72)	(4.10)	

		Acquisition Efficiency	Transfer Efficiency	
Device				
BOT	$\bar{x}$	2.50	5.17	3.84
	δ	(2.5)	(2.99)	
E3A	$\bar{x}$	3.50	6.83	5.06
	δ	(2.18)	(4.76)	
VIGS	$\bar{x}$	7.08	8.83	7.95
	δ	(3.36)	(5.15)	

TABLE 30. MEANS AND STANDARD DEVIATIONS* OF PAIRED RATER COMPARISONS FOR EACH TRAINING DEVICE - DEFT III

		Question					
		PD	LD	TA	RD	RLD	TT
Burst on Target Task 1							
Index Ammunition	$\bar{x}$ σ	0.00 (0.00)	0.00 (0.00)	24.64 (12.89)	-	0.00	23.75
Turn on Main Gun Switch	$\bar{x}$ σ	0.50 (0.50)	0.00 (0.00)	26.36 (9.59)	-	-	-
Announce Identified	$\bar{x}$ σ	0.00 (0.00)	0.00 (0.00)	22.44 (12.04)	-	-	-
Apply Lead (Simulated)	$\bar{x}$ σ	0.67 (0.47)	0.17 (0.17)	11.80 (9.11)	0.00	0.33 (0.24)	13.92 (6.30)
Lay Crosshair Leadline	$\bar{x}$ σ	0.00 (0.00)	0.00 (0.00)	11.44 (9.34)	-	0.00 (0.00)	16.94 (9.62)
Fire Main Gun	$\bar{x}$ σ	0.67 (0.47)	0.00 -	53.46 -	-	-	-
Task 2							
Sense Round	$\bar{x}$ σ	0.00 (0.00)	0.00 (0.00)	8.05 4.84	0.00	0.33 (0.24)	12.21 (6.87)
Announce Sensing & "BOT"	$\bar{x}$ σ	0.67 (0.47)	0.00 (0.00)	9.73 (7.03)	-	0.00 (0.00)	18.67 (6.60)
Relay to New Aiming Point	$\bar{x}$ σ	1.00 (0.58)	0.00 (0.00)	6.99 (4.46)	-	0.00 (0.00)	6.44 (3.82)
Fire Main Gun	$\bar{x}$ σ	0.67 (0.47)	0.00 -	48.91 -	-	-	-
E3A Task 1							
Ensure Power & Cooling Avail.	$\bar{x}$ σ	1.33 (0.94)	0.00 (0.00)	28.11 (13.90)	-	-	-
Turn on NCS Power on	$\bar{x}$ σ	1.83 (1.07)	0.00 (0.00)	28.23 (13.77)	-	0.17 (0.10)	24.94 (11.68)

* Standard deviations are provided when more than two raters supplied a rating.

Table 30 (Continued)

		PD	LD	TA	RD	RLD	TT
Turn Autopilot	$\bar{x}$	2.33	0.00	31.55	-	0.17	26.25
Off	σ	(1.37)	(0.00)	-			
Turn Probe	$\bar{x}$	2.33	0.00	31.55	-	0.17	27.50
Heaters Off	σ	(1.37)	(0.00)			No. D,G	No. D,G
Synchronize	$\bar{x}$	1.83	0.00	4.18	-	0.11	22.50
Horizontal	σ	(1.07)	(0.00)	(1.52)		(0.08)	(8.81)
Situation Indicators							
INS-1 & INS-2	$\bar{x}$	1.83	0.00	14.97	-	0.11	25.00
to Align Mode	σ	(1.07)	(0.00)	(6.24)		(0.00)	(11.90)
Test UDC	$\bar{x}$	1.83	0.17	14.85	-	0.11	24.17
Display & Lamps	σ	(1.07)	(0.17)	(5.87)		(0.08)	(9.80)
						No. G	No. G
Detect	$\bar{x}$	2.17	0.00	37.33	-	-	-
Fault 10	σ	(1.57)	(0.00)	(13.56)			
<u>Task 2</u>							
CDUs	$\bar{x}$	2.00	0.00	24.97	-	0.11	15.08
	σ	(1.41)	(0.00)	(12.23)		(0.08)	(7.07)
Sim. Restart,	$\bar{x}$	2.50	0.00	39.73	-	0.33	19.90
Perform	σ	(1.50)	-			(0.24)	(16.07)
Checkout							
INUs	$\bar{x}$	2.00	0.00	24.97	-	0.00	15.08
	σ	(1.41)	(0.00)	(12.23)		(0.00)	(7.07)
Sim. Restart,	$\bar{x}$	2.50	0.00	48.73	-	0.36	19.90
Perform	σ	(1.50)	-			(0.26)	(16.07)
Checkout							
Check 115 VAC	$\bar{x}$	1.50	0.17	22.70	-	0.78	9.25
Power	σ	(1.50)	(0.17)	(11.59)		(0.44)	(4.81)
Sim. Continuity	$\bar{x}$	2.00	0.22	27.64	-	0.33	13.00
Check, Check	σ	(1.41)	(0.16)	(10.48)		(0.14)	(5.70)
Wiring Continuity							
Sim. Replace.	$\bar{x}$	0.50	0.00	29.27	-	0.00	19.50
of Capacitor,	σ	(0.50)	(0.00)	(10.72)			
Replace Shorted Capacitor							

Table 30 (Continued)

		PD	LD	TA	RD	RLD	TT
Sim. Restart,	$\bar{x}$	2.50	0.00	48.73	-	0.33	25.17
Perform	σ	(1.50)				(0.24)	(17.80)
Checkout							
VIGS							
Task 1							
Index	$\bar{x}$	0.50	-	-	-	-	-
Ammunition	σ	(0.50)					
Turn on Main	$\bar{x}$	0.50	-	-	-	-	-
Gun Switch	σ	(0.50)					
Announce	$\bar{x}$	0.67	0.00	27.36	-	-	-
IDENTIFIED	σ	(0.47)					
Apply	$\bar{x}$	0.00	0.00	13.65	-	0.33	10.63
Lead	σ	(0.00)	(0.00)	(8.64)		(0.24)	(4.83)
Lay Crosshair	$\bar{x}$	0.50	0.00	11.77	-	0.00	6.98
Leadline	σ	(0.50)	(0.00)	(5.41)		(0.00)	(4.72)
Fire Main	$\bar{x}$	0.50	-	-	-	-	-
Gun	σ	(0.50)					
Task 2							
Sense	$\bar{x}$	0.67	0.00	18.76	-	0.00	10.46
Round	σ	(0.47)	(0.00)	(10.28)		(0.00)	(5.02)
Announce	$\bar{x}$	0.00	0.00	24.86	-	0.00	14.67
Sensing & "BOT"	σ	(0.00)	(0.00)	(11.59)		(0.00)	(5.35)
Relay to New	$\bar{x}$	1.17	0.22	18.44	-	0.00	12.08
Aiming Point	σ	(0.69)	(0.16)	(10.70)		(0.00)	(5.76)
Fire Main	$\bar{x}$	0.50	-	-	-	-	-
Gun	σ	(0.50)					

Summary

Based on the analyses presented in this report, a number of recommendations can be made regarding modifications of DEFT:

1. The expected distribution of summary index scores is too large to provide for meaningful interpretations of DEFT output, unless various assumptions are made regarding the expected distributions of input variables in the real world. All of the assumptions we made are defensible (e.g., a training device will not be built that addresses no performance deficit, etc.); however, a different set of assumptions would result in different critical values for inter-device comparisons.

2. The major contributors to output variance are the two Efficiency scales. To reduce this problem, it is recommended that some transform (e.g., square root) be used.

3. It is recommended that two additional scales be added to the DEFT II analyses. These scales would assess the proportion of required skills and knowledge contained in the training device requirement and the operational performance objective that the trainees do not possess.

4. It is imperative that when more than one rater applies DEFT to the evaluation of a device, the raters agree on their assumptions regarding the device, trainee population, device utilization, and the meanings of the various DEFT scales prior to conducting analyses.

Based on these results, recommendations 2 and 3 above have been implemented in the most recent DEFT programs. Presumably, the remaining recommendations would be implemented by DEFT users.

REFERENCES

Rose, A.R. & Wheaton, G.R. (1984a) **Forecasting device effectiveness: I. Issues.** Technical Report. Washington, DC: American Institutes for Research.

Rose, A.R. & Wheaton, G.R. (1984b) **Forecasting device effectiveness: II. Procedures.** Technical Report. Washington, DC: American Institutes for Research.

Harris, J.H., Ford, P. (HumRRO), Tufano, D., & Wiggs, C. (ARI). (1983) **Application of transfer forecast methods to armor training devices.** Alexandria, VA: Human Resources Research Organization.